Multi-person Pose Understanding in Complex Physical Interactions

Wen Guo

▶ To cite this version:

Wen Guo. Multi-person Pose Understanding in Complex Physical Interactions. Artificial Intelligence [cs.AI]. Inria - Research Centre Grenoble – Rhône-Alpes; l'Université Grenoble Alpes, 2023. English. ⟨NNT : ⟩. ⟨tel-04387188⟩

HAL Id: tel-04387188

https://hal.science/tel-04387188v1

Submitted on 11 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright - All rights reserved

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire Jean Kuntzmann

Interactions Physiques Complexes et Compréhension de la Pose Multi-personnes

Multi-person Pose Understanding in Complex Physical Interactions

Présentée par :

Wen GUO

Direction de thèse :

Xavier ALAMEDA-PINEDA Chargé de recherche HDR, INRIA de l'UGA	Directeur de thèse
Francesc MORENO-NOGUER Senior scientist, IRI	Co-directeur de thèse

Rapporteurs :

QIN JIN
Full professor, Renmin University of China
GERARD PONS-MOLL
Professeur, Eberhard Karls Universität Tübingen

Thèse soutenue publiquement le **12 juin 2023**, devant le jury composé de :

XAVIER ALAMEDA PINEDA Chargé de recherche HDR, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse
FRANCESC MORENO-NOGUER Senior scientist, IRI	Co-directeur de thèse
QIN JIN Full professor, Renmin University of China	Rapporteure
GERARD PONS-MOLL Professeur, Eberhard Karls Universität Tübingen	Rapporteur
SIYU TANG Professeur assistant, Ecole Polytechnique Fédérale de Zurich	Examinatrice
DIDIER SCHWAB Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Président

Abstract

Understanding the pose and motion of humans in 3D space has undergone enormous progress in recent years. The majority of studies in multi-person scenarios view individuals as separate instances, neglecting the importance of interaction information. However, in numerous everyday contexts, people always engage with one another, and it is essential to consider pose instances jointly as the pose of an individual is influenced by the poses of their interactees. In this context, this thesis aims to develop deep learning techniques to understand human pose and motion in complex interactions and leverage interaction information to enhance the performance of human pose estimation and human motion prediction. Our investigation encompasses modeling and learning interaction information, leveraging this information to refine human pose and motion estimation, and addressing the issue of insufficient 3D interacting human datasets. To overcome these challenges, we undertake the following steps: (1) we verified the feasibility of considering person interaction on the task of 3D human pose estimation from a single RGB image, by modeling and learning the interaction information with a deep network (PI-Net) on publicly available datasets (published at IEEE WACV2021); (2) we collected and released a new dataset for extreme interacting poses (ExPI dataset) to study person interaction (published at IEEE CVPR2022); (3) observing poses as temporal sequences, we study the task of collaborative motion prediction and propose a model with cross-interaction attention (XIA), using interaction information as guidance to improve multi-person motion prediction (published at IEEE CVPR2022); (4) rethinking the task of human motion prediction, we further propose a simple and effective baseline model for human motion prediction (published at IEEE WACV2023), which works not only on single-person motion prediction but also on multi-person scenarios. The code of our works is publicly available.

Résumé

La compréhension de la pose et du mouvement des humains dans l'espace tri-dimensionel a connu d'énormes progrès ces dernières années. Cependant, la majorité des études sur les scénarios multi-personnes considèrent les individus comme des instances distinctes, négligeant l'importance des informations d'interaction. Pourtant, au quotidien, les personnes interagissent beaucoup les unes avec les autres. Cette thèse vise donc à développer des techniques d'apprentissage profond pour comprendre la pose et le mouvement humain dans des scénario d'interactions complexes, en exploitant ces informations d'interaction pour améliorer les performances. Pour surmonter ces défis, nous avons entrepris les étapes suivantes : (1) Nous vérifions la possibilité de considérer l'interaction entre personnes pour l'estimation de la pose humaine en 3D à partir d'une seule image RVB, en modélisant et apprenant les informations d'interaction avec un réseau profond (PI-Net) sur des ensembles de données publiques (publié à IEEE WACV2021). (2) Nous collectons et publions un nouvel ensemble de données pour les poses avec interaction extrêmes (ExPI dataset) pour étudier l'interaction entre les personnes (publié à IEEE CVPR2022). (3) En observant les poses comme des séquences temporelles, nous étudions la prédiction de mouvement collaboratif et proposons un modèle basé sur un réseau d'attention pour l'interaction croisée (XIA), utilisant les informations d'interaction comme guide pour améliorer la prédiction de mouvement multi-personnes (publié à IEEE CVPR2022). (4) En repensant la tâche de prédiction de mouvement humain, nous proposons un modèle de base simple et efficace pour la prédiction de mouvement humain basé sur le perception mutli-couches (MLP), qui fonctionne non seulement pour la prédiction du mouvement d'une seule personne, mais aussi sur des scénarios multi-personnes (publié à IEEE WACV2023). Le code de nos travaux est disponible publiquement.

ACKNOWLEDGMENT

First and foremost I would like to express my sincere gratitude to my supervisors Prof. Xavier Alameda-Pineda and Prof. Francesc Moreno-Noguer. First thanks for accepting me as a Ph.D. student three years ago, and for all the help and guidance which helped me open the door to the fantastic world of research. Besides the valuable guidance on technical and research parts, I would also like to thank Xavi for all the kind advice and support during my whole journey as a Ph.D. student, and I would like to thank Francesc for all the days and nights ”fighting” together with us online until the last minute of deadlines. I feel really lucky to have you as my supervisors.

Besides, I would like to thank the help and guidance of other professors I worked with: Prof. Vincent Lepetit, Prof. Lauren Girin, and Prof. Simon Leglaive. And I would also like to thank my other collaborators: Yuming Du, Dr. Xi Shen, Xiaoyu Bie, and Enric Corona.

I would also like to express my sincere gratitude to my thesis committee members, Prof. Qin Jin, Prof. Gerard Pons-Moll, Prof. Siyu Tang, and Prof. Didier Schwab, for agreeing to review this work and participate in my defense.

I would like to thank all my other colleagues in the Robotlearn/Perception team, and also in the IRI team: Prof. Radu Horaud, Soraya, Zhiqi, Louis, Anand, Yihong, Xiaoyu Lin, Gaétan, Guillaume, Alex, Chris, Mostafa, Xiaofei, Sylvain, Nicolas, Ginger, Jianxiong, Ruijie. I truly enjoyed the time we spent together, no matter in person or online. Moreover, I would like to thank the staffs at Inria and IRI whose work directly or indirectly facilitated this thesis: especially Nathalie and Victor who greatly helped with the administrative aspects of my research. I would also thank the Kinovis platform at Inria Grenoble, Laurence Boissieux, and Julien Pansiot for their help.

I appreciate my parents and my sister for supporting and accompanying me going through this journey. And I would like to thank all my friends around during these three special years of the pandemic. Special thanks to my boyfriend, who is always there giving me support and encouragement and cheering me up, especially during these three special years far away from home.

CONTENTS

1 Introduction **11**

1.1 Motivation 12

1.2 Challenges and goals 15

1.3 Contributions 16

1.4 Manuscript Structure 19

1.5 Publications and Works Under Submission 20

2 Literature Review **21**

2.1 3D Pose Estimation 22

2.1.1 3D Single-person pose estimation 22

2.1.2 2D multi-person pose estimation 22

2.1.3 3D Multi-person pose estimation 23

2.1.4 Person interaction in human pose estimation 24

2.2 Human motion prediction 25

2.2.1 RNN-based human motion prediction 25

2.2.2 GCN-based human motion prediction 26

2.2.3 Attention-based human motion prediction 27

2.2.4 Person interaction in human motion prediction 27

3 Preliminary Research 29

3.1 Tasks and evaluation 30

3.2 Datasets 32

4 Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation 37

4.1 Introduction 39

4.2 Related Work 40

4.3 PI-Net for Multi-Person Pose Estimation 42

4.3.1 Pipeline of PI-Net 42

4.3.2 Interaction Order 44

4.3.3 Network Architecture 45

4.3.4 Implementation details 46

4.4 Experiments 46

4.4.1 Datasets 46

4.4.2 Baseline and Evaluation metrics 47

4.4.3 Main results 48

4.4.4 Ablation Study 49

4.5 Conclusion 55

5 Multi-Person Extreme Motion Prediction 57

5.1 Introduction 59

5.2 Related Work 61

5.3 Problem Formulation 63

5.4 The Extreme Pose Interaction Dataset 64

5.4.1 Dataset Collection and Post-processing. 64

5.4.2 Dataset Structure 65
5.4.3 Data Analysis 68
5.5 Method 70
5.5.1 Pipeline 71
5.5.2 Cross-Interaction Attention (XIA) 73
5.5.3 Pose Normalization 75
5.6 Experimental Evaluation 75
5.6.1 Splitting the ExPI Dataset 77
5.6.2 Evaluation Metrics 77
5.6.3 Implementation Details 78
5.6.4 Results and Discussion 79
5.7 Conclusion 82

6 A Simple Baseline for Human Motion Prediction 89

6.1 Introduction 90
6.2 Related Work 93
6.3 Our Approach: SIMLPE 94
6.3.1 Discrete Cosine Transform (DCT) 94
6.3.2 Network architecture 97
6.3.3 Losses 98
6.4 Experiments for single-person human motion prediction 99
6.4.1 Datasets and evaluation metric 99
6.4.2 Implementation details 101
6.4.3 Quantitative and qualitative results 102

6.4.4 Ablation study 104
6.5 Experiments for multi-person human motion prediction 106
6.5.1 Datasets and evaluation metric 106
6.5.2 Quantitative results 106
6.6 Conclusion 107

7 Conclusions and Future Work 113
7.1 Summary 114
7.2 Future research directions and discussions 115
7.2.1 Future follow-up works 115
7.2.2 A broader discussion within the domain of 3D human pose and motion understanding 116

List of Figures 123

List of Tables 128

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Computer vision strives to empower computers with the ability to better understand the world with visual input, thus understanding human behavior and interactions plays a significant role. In recent years, human understanding tasks in 3D space, such as 3D human pose and motion understanding, have garnered substantial research and industrial interest. This is largely due to the numerous potential applications across many different fields, including sports technology, physical therapy, medical diagnosis, robotic manipulation, autonomous navigation, the entertainment industry, virtual reality, etc. Advancements in 3D human pose and motion understanding can greatly enhance the accuracy and efficiency of these applications, leading to improved user experiences and outcomes.

Robotics. 3D human pose and motion understanding could offer numerous valuable applications in the field of robotics. One key aspect is enabling robots to comprehend and predict the motion of users, allowing them to provide appropriate responses or even physical reactions. This capability has wide-ranging applications in healthcare, elderly and child care, and intelligent services in public spaces such as hotels, banks, and shopping malls. For instance, if a patient requires assistance from a robot to walk to the restroom, the robot should be able to understand the patient's movement, predict their arrival, and provide precise support based on the detected body joints, such as the location of the hands, elbows, legs, etc. (Figure 1.1 (b)).

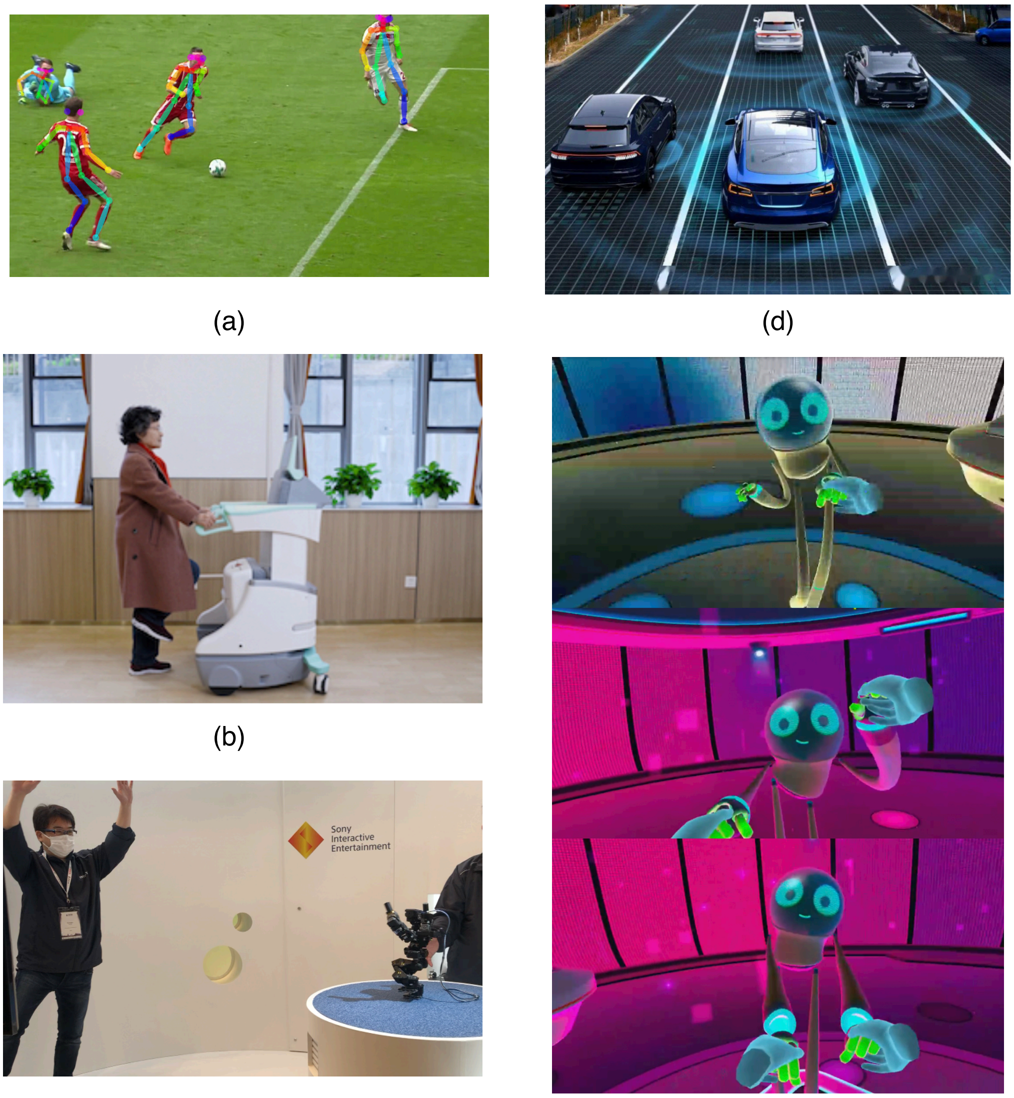
Another application involves robots understanding user poses and accurately replicating their movements (Figure 1.1 (c)). This functionality can enable people to remotely operate robots in hazardous environments, where direct human presence would be risky. For example, users could remain in safe areas while directing robots to work in dangerous factory settings, disaster relief operations, or even perform tasks outside space stations as astronauts.

Besides, both of the above aspects could be used in the entertainment industry, creating innovative and interactive experiences for users.

Autonomous driving. 3D Human pose and motion understanding could also help the development of autonomous driving (Figure 1.1 (d)). Specifically, understanding and predicting the trajectory and motion of the pedestrians could help the system to avoid possible accidents in advance, or give alerts to the human driver. By accurately analyzing pedestrian poses and motions in real-time, autonomous driving systems can proactively avoid potential hazards. For instance, an autonomous vehicle could adjust its speed or change lanes to prevent a collision with a pedestrian crossing the road. This technology facilitates safe and efficient navigation for vehicles in complex environments, enabling them to navigate effectively. In addition to enhancing autonomous vehicle safety, 3D human pose and motion understanding can improve the driving experience for human drivers. Advanced driver-assistance systems (ADAS) can provide timely alerts and warnings, aiding drivers in avoiding accidents and making better decisions. By fostering a deeper understanding of human behavior in traffic situations, we can develop more intelligent transportation systems that prioritize safety, efficiency, and user experience.

Sports science. Understanding and analyzing the poses and motions of athletes plays a significant role in sports science (Figure 1.1 (a)). By leveraging human pose and motion understanding techniques, real-time pose data of athletes can be easily recorded and analyzed. This could be helpful for coaches and athletes to identify areas for improvement, develop personalized training programs, and monitor progress over time, thus benefiting the advancement of sports technology and performance analysis. Furthermore, it also enhances live sports events for spectators: by providing detailed insights into athlete movements and biomechanics, viewers can better appreciate the skill, technique, and effort involved in various sports.

Virtual reality Based on the study of 3D human motion, multiple applications in virtual reality could be realized. For example, having only controllers on two hands and sensors on the head for the user which provides half-body keypoints, legs could be also generated in the virtual world to obtain a whole-body virtual avatar for the user. Also, based on the motion of the user, virtual characters could interact with and react to the user with reasonably generated motions, like dancing with the user (Figure 1.1 (e)). Such immersion and interaction enhance the virtual reality experience for users.



(a) (d)

(b)

(c) (e)

Figure 1.1: Applications of human understanding: (a) sport analyses; (b) a hospitalization robotic which takes care of patients; (c) an entertainment robotic which could follow the motion of the user; (d) autonomous driving; (f) virtual avatar dancing with the user. [2]

---

[2] These figure are from:
(b) https://news.xinmin.cn/2022/09/01/32224646.html
(c) IROS2022 Sony Interactive Entertainment, photo taken by the author.
(d) https://www.sohu.com/a/569062231_468661
(e) Starting demo from Oculus Meta Quest2, photo taken by the author.

1.2 CHALLENGES AND GOALS

Single 3D human pose and motion understanding has had fast developments in recent years, while in multi-person scenarios, the problem of modeling the relations between different people remains under-investigated. Previous works usually treat different people separately, disregarding the interactive information among them. These approaches are agnostic about the context information such as the existence of other people around [138, 122, 139, 118, 84, 164, 117, 17, 40]. However, according to previous psychophysics studies [127], this is not how our visual system understands the world: objects in the real world never occur in isolation but co-vary with each other, thus giving a rich source of contextual associations. A natural way to represent the context of a person is in terms of their relationship with other instances sharing the same environment. These contextual associations might include the interaction of a person with objects around or with other persons involved in the same scenario.

In this thesis, we focus on studying human pose and motion in the multi-person scenario with person interactions: when we human beings look at a multi-person scenario, especially when different persons are involved in the same activity, how do we understand the behavior of a person is highly related to the observation of other persons around or the persons involved in the same activity. Thus, we aim to develop machine learning techniques able to jointly estimate the full body pose or motion (i.e. a sequence of poses) of several people involved in the same complex physical interaction, seeking to explore the use of person-interaction information among different people to improve the understanding of human poses and motions. We focus on two tasks: we start by considering a single frame and look into the task of 3D human pose estimation from a single image. Based on publicly available datasets, we study how to improve the performance of 3D pose estimation with the help of interaction among different people in the same scene, thus verifying the feasibility of studying human interaction. There are two challenges for this problem: (1) How to model and learn the interaction information; (2) How to use the interaction information to improve the predicted poses.

Next, we progress from a single frame to a sequence of poses and shift our attention to the task of human motion prediction, aiming to forecast future pose sequences based on past observations. This task also presents two challenges: (1) Properly embedding and learning the interaction information between two interacting motion sequences, and (2) The lack of suitable multi-person interaction data for studying this problem. While investigating this issue, we encounter an additional challenge: In spite of complicated and heavy models, could a simple and light model have state-of-the-art performance for forecasting human motion? In response to these challenges, we have conducted a series of studies and experiments, where the main contributions are introduced in the section below.

1.3 CONTRIBUTIONS

Based on the challenges described above, we started our research by investigating human interaction in the context of multi-person monocular 3D pose estimation using publicly available datasets. In this study, we fed the initial pose along with its corresponding interacting poses into a recurrent network to refine the pose of the target individual. Our method proves its effectiveness on the public MuPoTS dataset, achieving a new state-of-the-art performance. This finding verifies the feasibility of incorporating human interaction in the process of understanding human poses.

In the dataset we used for the above studies, the interaction signal is weak. At that moment, existing publicly available 3D multi-person human pose datasets are either too small, not captured in real-world scenarios, or lack a significant number of highly interactive actions. This makes it challenging to learn person interaction using data with limited real interaction signals. Consequently, annotated data with complex interactions becomes crucial for studying human interactions. Since no such dataset is available, we have captured ExPI (Extreme Pose Interaction) dataset, a new lab-based person interaction dataset of 2 couples of professional dancers performing lindy-hop dancing actions containing 115 sequences with 30k frames and 60k instances with annotated 3D body poses and shapes.

We have carefully cleaned and checked the data manually to ensure the high quality of the data.

Furthermore, during our study of human pose estimation from RGB images, we discovered that modeling interaction within a single frame is challenging, whereas sequence signals provide clearer and less misleading information about human interactions. Consequently, we began investigating interactive human motion prediction using mocap sequences from the ExPI dataset. Leveraging the ExPI dataset, we advanced beyond existing 3D human motion prediction approaches by considering scenarios involving two people engaged in highly interactive activities. Traditional motion prediction formulations focus on a single individual, while to learn coupled motion dynamics, we introduced a novel mechanism that utilizes historical information from both people in an attention-like manner. This model was trained using our ExPI dataset. The results of the provided cross-interaction attention (XIA) mechanism demonstrate consistent improvements compared to baseline models that predict the motion of each person independently.

While investigating human motion prediction, we noticed that recent studies on single-person human motion prediction have increasingly focused on designing more complex architectures to improve performance. Although these state-of-the-art approaches yield impressive results, they depend on intricate deep learning architectures, such as Recurrent Neural Networks (RNNs), Transformers, or Graph Convolutional Networks (GCNs). These models typically necessitate multiple training stages and often involve more than 2 million parameters. We began by conducting two straightforward experiments: repeating the last frame of input observation to serve as output prediction and using a single FC layer to see how it performs. These experiments yielded reasonably good results, suggesting that a simpler network might be sufficient for this task. Consequently, we introduced a simple yet effective network for human motion prediction based on extensive experimentation. The proposed method comprises only fully connected layers, layer normalization, and transpose operations, with layer normalization being the sole non-linear operation. Despite having significantly fewer parameters, this method attains state-of-the-art performance on various benchmarks. When applied to the ExPI dataset, which features multi-person scenarios, our simple network (SiMLPe) achieves results comparable to state-of-the-art approaches with minimal adaptations.

To summarize, the contributions of this thesis are listed as below:

- We explore ways to leverage the interdependencies between individuals in a shared scenario to enhance current and potentially future deep networks for 3D monocular pose estimation. Our Pose Interacting Network (PI-Net) takes the initial pose estimates of a varying number of interactees and refines the pose of the person of interest using a recurrent architecture. By achieving state-of-the-art results on the MuPoTS dataset and producing strong qualitative results on the COCO dataset, we verify the effectiveness of incorporating interaction information to improve pose estimations. This work was presented at WACV2021, and the code is available at `https://github.com/GUO-W/PI-Net`.

- We collect and release the ExPI (Extreme Pose Interaction) dataset, a new lab-based person interaction dataset of professional dancers performing dancing actions annotated with challenging 3D body poses and shapes in highly interacted situations. The provided data enables studies of multiple human understanding tasks and especially studies of human interaction. The data is available at `https://zenodo.org/record/5578329#.Y8Gw8OzP23K`.

- Transitioning from single-image to sequence motion, we investigate the task of collaborative human motion prediction when dealing with humans performing collaborative tasks. We seek to forecast the future movement of two individuals who interact with each other, based on their past skeletal sequences. To achieve this, we introduce a new cross-interaction attention mechanism that leverages historical data from both individuals and models the relations between their pose sequences. We extensively evaluate our cross-interaction network on the ExPI dataset, demonstrating that it consistently outperforms state-of-the-art single-person motion prediction methods for both short- and long-term predictions. This work, along with the ExPI dataset, was presented at CVPR2022, and the code of this work is available at https://github.com/GUO-W/MultiMotion.

- We proposed a simple but strong baseline for motion prediction. Our experiments demonstrate that a lightweight network consisting of multi-layer perceptrons (MLPs) with just 0.14 million parameters can outperform the state-of-the-art methods for motion forecasting, combining with a set of standard techniques including Discrete Cosine Transform (DCT), residual joint displacement prediction, and velocity optimization using an auxiliary loss. The proposed model, siMLPe, consistently outperforms all other approaches on Human3.6M, AMASS, 3DPW, and ExPI datasets for single- and multi-person motion prediction. This simple method has also brought some insights into other motion-understanding tasks such as human motion generation. This work was presented at WACV2023, and the code of this work is available at https://github.com/dulucas/siMLPe.

1.4 MANUSCRIPT STRUCTURE

This manuscript is organized as below: Chapter 2 proposes a general literature review of the previous related works; Chapter 3 formulates the tasks of human pose estimation and human motion prediction concerned in this thesis, introduces the evaluation metrics and the datasets used; Chapter 4 illustrates in detail the proposed model PI-Net in monocular 3D human pose estimation; Chapter 5 introduces a new dataset ExPI, and details the proposed XIA model for collaborative motion prediction, evaluated following our carefully designed evaluation protocols for this new task; Chapter 6 proposes a single and effective baseline model for human motion prediction, which leads to a rethinking of this task; Chapter 7 recalls our contributions and discusses the potential future research directions in this domain.

1.5 PUBLICATIONS AND WORKS UNDER SUBMISSION

This section is the list of papers published or submitted during my Ph.D. The following three works are discussed in this manuscript:

- [61] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023*
- [60] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.*
- [62] Wen Guo, Enric Corona, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Pi-net: Pose interacting network for multi-person monocular 3d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021.*

There are also some other works or preprints which are not discussed in this thesis:

- [19] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. *arXiv preprint arXiv:2204.01565, 2022.*
- [59] Wen Guo. Multi-person pose estimation in complex physical interactions. In *Proceedings of the 28th ACM International Conference on Multimedia (Doctoral Symposium), 2020.*
- [44] Yuming Du, Wen Guo, Yang Xiao, and Vincent Lepetit. 1st Place Solution for the UVO Challenge on Image-based Open-World Segmentation 2021. *arXiv preprint arXiv:2110.10239.*

CHAPTER 2

LITERATURE REVIEW

2.1 3D POSE ESTIMATION

2.1.1 3D SINGLE-PERSON POSE ESTIMATION

Single-person 3D pose estimation using deep learning methods can be divided into two different strategies. The first strategy is to directly learn the mapping from image features to 3D poses [115, 98, 132, 136, 35]. Li *et al.* [98] propose a joint model for body part detectors and pose regression. Pavlakos *et al.* [132] introduce a U-Net architecture to recover joint-wise 3D heatmaps. Sun *et al.* [149] use a bone-based representation that enforces human pose structure for regression, and in [150], a differentiable soft-argmax operation is used for efficiently training an hourglass network.

The second strategy focuses on recovering 3D human pose from 2D image features by using models that enforce consistency between 3D predicted poses and 2D observations. For example, Bogo *et al.* [20] fits a human body parametric model by minimizing the distance between the projection of the 3D estimation and the 2D predicted joints. Moreno-Noguer [123] proposes to infer 3D pose via distance matrix regression, while Yang et al. [173] use an adversarial approach to ensure that estimated poses are anthropomorphic.

2.1.2 2D MULTI-PERSON POSE ESTIMATION

Multi-person pose estimation in 2D is well explored. There are two main approaches to multi-person pose estimation: top-down [171, 99, 148, 32] and bottom-up models [30, 29, 135, 126].

In the first approach, a human detector is used to estimate the bounding boxes containing the person in an image. Each detected area is then cropped and fed into the pose estimation network to estimate the poses of each person one by one. This pipeline requires a separate human detection step, and the pose estimation is performed independently on each detected person.

In contrast, the second approach follows a different pipeline. A model is used to estimate all human body keypoints in an image. These keypoints are then grouped into each person using clustering techniques. This pipeline does not require a separate human detection step, and the pose estimation is performed on all the keypoints in the image at once. For instance, Cao *et al.* [30, 29] proposed a famous bottom-up approach that uses Part Affinity Fields to group joints of different people in real time. The bottom-up approaches could be used as a backbone for lifting 2D joints to 3D in subsequent stages [40, 131]. The bottom-up approaches are efficient as they allow for the processing of multiple individuals in the same image without requiring a separate human detection stage.

Bottom-up methods generally outperform one-stage bottom-up methods. For example, Xiao *et al.* [171] developed a simple but effective baseline by using ResNet [67] as an encoder and several deconvolutional layers as a decoder. Sun*et al.* [148] maintained richer semantic information by connecting high-to-low resolution convolution streams in parallel. Chen *et al.* [32] improved the accuracy of hard keypoints in the initial estimates by using a cascade pyramid network.

Besides the effectiveness of these well-designed methods, another reason for the fast development of multi-person 2D pose estimation is the release of huge-scale datasets such as MSCOCO dataset [101], AI-Challenger Dataset [170], CrowdPose Dataset [93], etc, which contains large scale datasets for multi-person under various and complex scenes.

2.1.3 3D MULTI-PERSON POSE ESTIMATION

Similar to their 2D counterparts introduced above, 3D multi-person poses estimation methods could also be split into two categories: the top-down approaches [138, 139, 122, 84, 169], and the bottom-up approaches[118, 117, 176].

Top-down methods typically start with a human detection step to estimate the bounding boxes containing each person in the image. Each detected area is then cropped and fed into the pose estimation network, which estimates the 3D joint locations. For example, Rogez *et al.* [138, 139] classifies the 2D bounding boxes of each person into one clustered anchor pose, and than this prior anchor pose is refined in a coarse-to-fine manner. Moon *et* *al.* [122] proposed an architecture that simultaneously predicts the absolute position of the root joint in 3D space and reconstructs the relative 3D body pose of multiple people.

In contrast, bottom-up methods do not rely on human detection and instead estimate all 3D joint locations in the image first. Then, clustering techniques are used to group the joint locations belonging to each individual person. For example, Mehta *et al.* [118, 117] first estimates three occlusion-robust location-maps [120], and then modeles the association between body keypoints by Part Affinity Fields [30]. Zanfir *et al.* [176] formalizes the problem of localizing and grouping people as a binary linear integer program and solves it by integrating a limb scoring model.

2.1.4 PERSON INTERACTION IN HUMAN POSE ESTIMATION

The above works presented do not consider interactions between different people when studying the multi-person problems. Recently, there has been an increased focus on incorporating contextual information in 3D pose estimation methods by integrating scene constraints[175] or considering the depth-order to resolve the overlapping problem [77, 92]. Jiang *et al.* [77] propose a depth ordering-aware loss that takes into account the occlusion relationship and interpenetration of people in multi-person scenarios. This approach improves the accuracy of the estimated joint locations by enforcing consistency with the relative depth ordering of the people in the scene. Similarly, Li *et al.* [92] divide human relations into different levels and define three corresponding losses to ensure that the orders of different people or different joints are correct or not. This approach takes into account the semantic relationships between different parts of the human body and helps improve the accuracy of the estimated poses. While these approaches consider contextual information, they do not explore the interaction relations between different people in the same activity. More recently, Fieraru *et al.* [47] proposed a new dataset of human interactions with several daily interaction scenarios and proposed a framework based on contact detection over model surface regions. The dataset was recently released in 2022, and it provides a valuable resource for developing and evaluating methods that can take into account the interactions between people in daily interaction scenarios.

2.2 HUMAN MOTION PREDICTION

Human motion prediction is formulated as a sequence-to-sequence task, where past observed motion is taken as input to predict the future motion sequence. Traditional methods for human motion prediction have utilized nonlinear Markov models [87], Gaussian Process dynamical models [165], and Restricted Boltzmann Machines [151]. While effective for predicting simple motions, these approaches struggle with complex and long-term motion prediction [49].

Recently, with the rise of deep learning, human motion prediction has seen significant advancements through the use of deep networks, such as Recurrent Neural Networks (RNNs) [49, 76, 114, 103, 33], Graph Convolutional Networks (GCNs) [113, 111, 60, 107, 41, 96, 94] and Transformers [111, 5, 26]. In this section, we will introduce the representative methods using these deep networks for human motion prediction.

2.2.1 RNN-BASED HUMAN MOTION PREDICTION

Due to the inherent sequential structure of human motion, some works address 3D human motion prediction by recurrent models. Fragkiadaki *et al.* [49] propose an encoder-decoder framework to embed human poses and an LSTM to update the latent space and predict future motion. Jain *et al.* [76] incorporated the semantic similarity between different body parts manually and utilized structural RNNs to propagate this information. However, these methods suffer from discontinuity and are only trained on action-specific models, meaning that a single model is trained for each specific action. Martinez *et al.* [114] proposed training a single model for multiple actions, instead of action-specific models, allowing the network to leverage regularities across different actions in large-scale datasets. This approach has been widely adopted by subsequent works. Additionally, they introduced a residual connection to model velocities instead of absolute values, resulting in smoother predictions.

Despite their effectiveness, the aforementioned methods suffer from inherent limitations of RNNs. Firstly, as sequential models, RNNs are difficult to parallelize during training and inference. Secondly, memory constraints prevent RNNs from exploring information from farther frames. To address these issues, some works have used RNN variants [103, 33], sliding windows [24, 25], convolutional models [68, 89], or adversarial training [56]. However, these methods often result in complicated networks with a large number of parameters.

2.2.2 GCN-BASED HUMAN MOTION PREDICTION

To better encode the spatial connectivity of human joints, the most recent works usually represent human poses as graphs and utilize Graph Convolutional Networks (GCNs) [147, 79] for the task of human motion prediction.

GCNs were first exploited for human motion prediction in Mao *et al.* [113]. They use a stack of blocks consisting of GCNs, nonlinear activation, and batch normalization to encode the spatial dependencies, and leverage discrete cosine transform (DCT) to encode temporal information. This work inspired most of the GCN-based motion prediction methods in recent years. Based on [113], Mao *et al.* [111] further improved the temporal encoding by cutting the past observations into several sub-sequences and adding an attention mechanism to find similar previous motion sub-sequences in the past with the current observations. Thus, the future sequence is computed as a weighted sum of observed sub-sequences. Then, a GCN-based predictor, the same as in [113], is used to encode the spatial dependencies. In contrast to using DCT transformation to encode input sequences, [86] utilized a multi-scale temporal input embedding approach, incorporating various-sized convolutional layers for different input sizes to enable different receptive fields in the temporal domain. Ma *et al.* [107] proposed two variants of GCNs to extract spatial and temporal features. They developed a multi-stage structure, where each stage contains an encoder and a decoder. During training, the model is trained with intermediate supervision to learn progressively refined prediction. Additionally, [41, 96, 94] extended the graph representation of human pose to a multi-scale version across the abstraction levels of human pose.

2.2.3 ATTENTION-BASED HUMAN MOTION PREDICTION

With the rise of Transformers [160], recent works [111, 5, 26] have attempted to use attention mechanisms for human motion prediction. Mao *et al.* [111] utilized attention to identify temporal relations. Aksan *et al.* [5] employed a combination of "spatial attention" and "temporal attention" to map both the temporal dependencies and the pairwise relations between joints. Cai *et al.* [26] used a transformer-based architecture with a progressive-decoding strategy to predict the DCT coefficients of target joints progressively based on the kinematic tree. To guide predictions, they constructed a memory-based dictionary to preserve the global motion patterns in the training data.

2.2.4 PERSON INTERACTION IN HUMAN MOTION PREDICTION

Modeling interactions and the contextual information has been proven to be effective in the topic of 3D human pose estimation [92, 65, 175, 167, 77, 61]. Contextual information has also been shown to be beneficial in predicting human path trajectories. For this purpose, recent works explore the use of multi-agent context with social pooling mechanisms [6], tree-based role alignment [46], soft attention mechanisms [162] and graph attention networks [74, 83, 90]. Unlike the trajectory forecasting problem that focuses on a single center point, motion prediction aims at predicting the dynamics of the whole human skeleton. Incorporating contextual information in such a situation is still much unexplored. Corona *et al.* [36] expand the use of contextual information into motion prediction with a semantic-graph model, but only weak human-to-human or human-to-object correlations are modeled. Cao *et al.*[28] involve scene context information into the motion prediction framework, but without human-to-human interaction. Adeli *et al.*[1, 2] develop a social context-aware motion prediction framework, where interactions between humans and objects are modeled either with a social pooling [1] or with a graph attention network [2]. However, they only study in 2D space [9] or with weak human interactions [164]. Since in this dataset [164], most of the actions involve weak interactions like shaking hands or walking together.

CHAPTER 3

PRELIMINARY RESEARCH

3.1 TASKS AND EVALUATION

This section introduces the definition and common evaluation method of the two tasks studied in this thesis: 3D human pose estimation and 3D motion prediction.

3D human pose estimation from single RGB input focuses on only one frame, taking a RGB image as input and predict the 3D poses from the 2D image at that instance. A predicted poses $P \in \mathbb{R}^{N \times J \times 3}$ is usually represented as 3D joint coordinates where N is the number of person in the frame and J is the number of joints representing the pose of each person. 3D motion prediction looks into a sequence of human poses, predicting future poses based on past observations. The poses could be represented in multiple parameterizations such as i) angle-joints, Euler angle, rotation metric, exponential maps etc., or ii) directly 3D joint coordinates. The former kind of representation needs a kinematic tree for the pose, representing the structure of the pose. The latter represents the joint locations directly but thus might be possible to suffer from the error of bone lengths of a pose. As we are focusing on multi-person scenarios, building up a kinematic tree to represent multi-person together is problematic, thus we just use the 3D joint coordinates to represent the poses.

Evaluation. The Mean Per Joint Position Error (MPJPE) is the most common metric used to evaluate the accuracy of 3D joint positions in pose estimation and motion prediction tasks. It measures the average Euclidean distance between the predicted joint positions and the corresponding ground-truth positions. The MPJPE is calculated as follows:

$$\text{MPJPE}(P, G) = \frac{1}{J} \sum_{j=1}^{J} \| P_j - G_j \|_2 , \tag{3.1}$$

where J is the number of joints, Pj is the estimated position of joint j, and G_j is the ground-truth position of joint j. The MPJPE is the average L2-norm across different joints between the prediction and ground-truth.

Rigid alignment When evaluating poses, in addition to directly comparing the predicted results with the ground truth, another common evaluation approach is to assess the poses after performing the rigid alignment. This evaluation ignores the global rotation and translation but just focuses on the error of the poses between the predicted and the groundtruth. To make the representation simple, we use P to represent the points of the predicted pose, and G to represent the points of the groundtruth, than the rotation and translation matrix for aligning P based on G is calculated by the following steps in Algorithm 1.

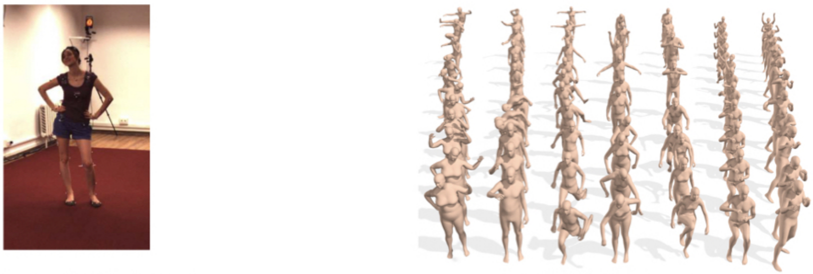
```
Require: $P, G$;
1: Compute the centers of $P$ and $G$:
       $P_c = \frac{1}{N}\sum_{i=1}^{N} P_i, G_c = \frac{1}{N}\sum_{i=1}^{N} G_i$        (3.2)
2: Set the origins to the centers:
       $P_i^{orig} = P_i - P_c, G_i^{orig} = G_i - G_c$        (3.3)
3: Singular value decomposition (SVD) of the correlation matrix:
       $C = \sum (P^{orig} G^{orig^T}) = U \Sigma V^T$        (3.4)
4: Estimate the rotation between the two point sets $P^{orig}$ and $G^{orig}$:
       $R = UV^T, T = P_c - RG_c$        (3.5)
5: return $\hat{P} = R\hat{P} + T$;
```

Algorithm 1: Rigid Alignment: Finding absolute orientation for 2 groups of points

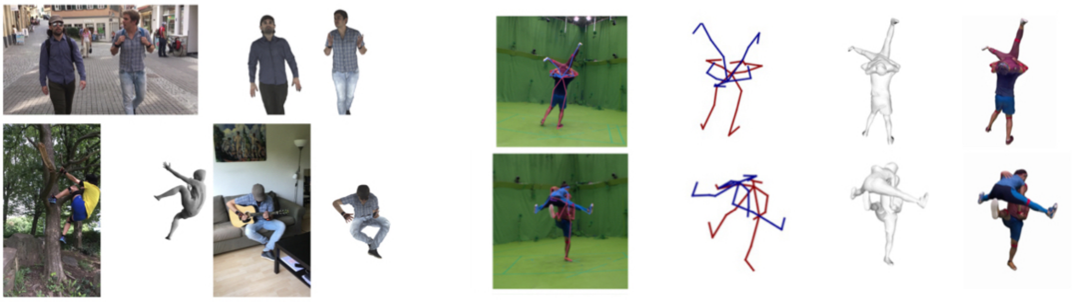
Then Aligned MPJPE could be calculated as

$$\text{MPJPE}_{align}(P, G) = \frac{1}{J}\sum_{j=1}^{J} \left\| \hat{P}_j - G_j \right\|_2 . \tag{3.6}$$

(a) MuCo3DHP dataset (b) MuPoTS-3D dataset (c) COCO dataset



(d) Human3.6M dataset (e) AMASS dataset



(f) 3DPW dataset (g) ExPI dataset

Figure 3.1: Datasets used in this thesis. (a, b, c) are datasets we use for human pose estimation in Chapter 4, (d, e, f) are datasets we use for human motion prediction in Chapter 6, and (g) is a dataset we collect and present in Chapter 5.

3.2 DATASETS

This section presents the datasets used in this thesis. Figure 3.1 shows some examples of these datasets.

MuCo3DHP dataset and MuPoTS-3D evaluation set. Multi-person Composited 3D Human Pose (MuCo3DHP) dataset and Multi-person Pose Test Set in 3D (MuPoTS-3D) test set was initially introduced by Mehta et al. [118]. MuCo3DHP dataset is a compositing-based training set based on MPI-INF-3DHP. It takes 1 to 4 subjects from the MPI-INF-3DHP [116] single-person 3D pose dataset, which contains 8 subjects with 2 clothing sets each and is captured by 14 cameras, and put them together as multi-person scenes. MuPoTS-3D evaluation set consists of 8320 images, each containing 2 or 3 people performing a common activity such as talking, shaking hands, or engaging in sports. The images were captured in 20 real scenes, which include 5 indoor scenes and 15 outdoor scenes, and contain a total of $23k$ instances of human poses. Each scene consists of several hundred frames extracted from a video, providing a diverse set of poses and activities to evaluate the performance of pose estimation methods.

The annotations of MuCO and MuPoTS-3D datasets are in COCO format, and these two datasets provides 2D image coordinates and 3D camera coordinates for each body joint. MuCO dataset is typically used as train data for the task of multi-person 3D human pose estimation, always combining with MuPoTS-3D dataset which serves as test data in the task of multi-person 3D human pose estimation.

COCO dataset. Common Objects in Context (COCO) dataset [101] is a large-scale object detection, segmentation, and captioning dataset which contains more than 200,000 images and 250,000 person instances labeled with key points. Although the COCO val2017 dataset only provides 2D ground-truth labels, it contains challenging scenes with a large number of people performing diverse actions, making it a valuable resource for evaluating the performance of 2D and 3D multi-person pose estimation methods. In particular, we use images from COCO val2017 subset for qualitative evaluation in this thesis.

Human3.6M dataset. Human3.6M contains 3.6 million 3D human poses and corresponding images, which includes 7 actors performing 15 actions with publicly released 3D labels. 32 joints are labeled for each pose. The datasets propose video from 4 calibrated cameras, 3D joint positions and joint angles from a motion capture system, 3D laser scans of the actors, and also person bounding boxes.

AMASS dataset. Archive of Motion Capture as Surface Shapes (AMASS) [109] is a large database of human motion that unifies multiple marker-based motion capture

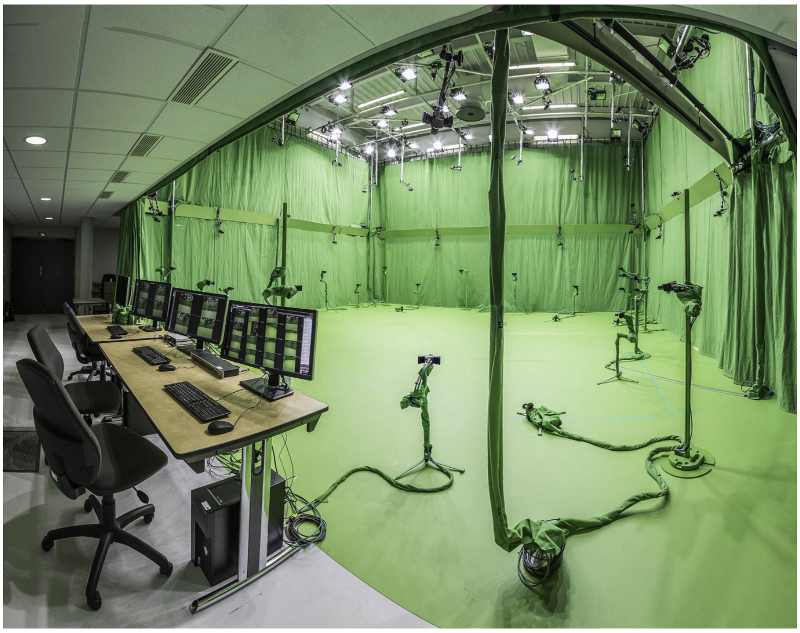


Figure 3.2: Inria Kinovis-platform.

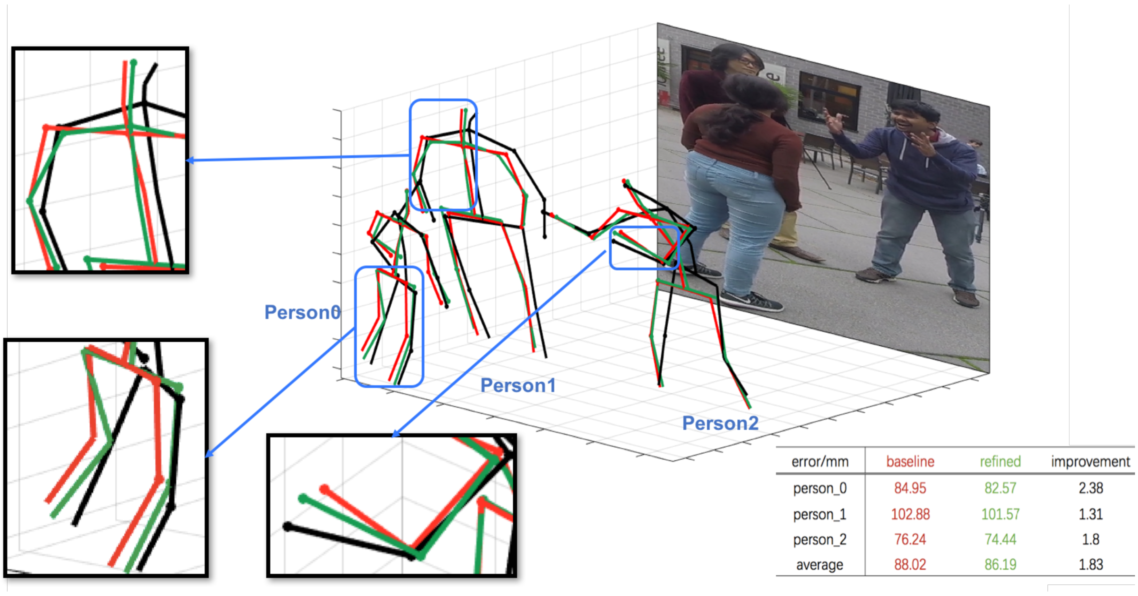
datasets [48, 109, 12, 85, 155, 157, 21, 106, 51, 31, 143, 110, 104, 124, 3, 158, 71, 156] and unifies these datasets within a common framework and parameterization SMPL [105]. It includes totally more than 40 hours of motion data with more than 300 subjects and more than 11000 motions.

3DPW dataset. 3D Poses in the Wild (3DPW) dataset [163] is an in-the-wild dataset with accurate 3D poses for evaluation, which includes video footage taken from a moving phone camera. This dataset contains 60 sequences with 2D and 3D annotations as well as 3D body scans and 3D people models.

ExPI dataset. Extreme Pose Interaction (ExPI) dataset is a dataset presented by Guo et al. [60]. It is a person interaction dataset of acrobatics, acquired by recording the performance of 2 couple of actors under 67 to 68 different viewpoints, for exploiting 3D human interaction and studying human behaviors such as human pose estimation, motion prediction, human shape estimation, etc. The dataset includes 29.6k frames with 115 sequences and 16 different actions, containing over 59.1k instances with annotated 3D body joints and shapes. The data were collected by the Kinovis-platform of INRIA which is a platform for capturing the moving human body, see Figure 3.2. The practical acquisition space of this platform is approximately $40m^2$. It has 2 acquisition systems: A 68-color camera (4MPixels) system that provides full shape and appearance information with 3D textured meshes, and a standard Motion capture (Mocap) system composed of 20 cameras that provides infrared reflective marker trajectories. The data was recorded by 18 markers on each actor. For more details about the dataset, please refer to Chapter 5.

CHAPTER 4

POSE INTERACTING NETWORK FOR MULTI-PERSON MONOCULAR 3D POSE ESTIMATION



error/mm	baseline	refined	improvement
person_0	84.95	82.57	2.38
person_1	102.88	101.57	1.31
person_2	76.24	74.44	1.8
average	88.02	86.19	1.83

Figure 4.1: **PI-Net peformance.** An example of test results on MuPoTS dataset. Poses refined by PI-Net (in green) are closer to the ground truth (in black) than the baseline (in red). We zoom in to several parts to clearly appreciate the difference. The error before and after PI-Net refinement for each person is shown in the table. The average 3D joint error for this example is reduced from 88.02 mm to 86.19 mm.

This chapter studies person interaction in the task of multi-person monocular 3D pose estimation, which takes a single image as input and aims to predict the 3D joint locations of multiple persons in the image. Previous literature addressed this problem satisfactorily, while in these works different persons are usually treated as independent pose instances to estimate. In this chapter, we investigate how to focus on the interaction between different people to improve the prediction results obtained by a prior 3D pose estimator by our proposed pose interacting network, PI-Net, which processes the initial estimates by a recurrent network and refines the poses. We train the model on publicly released datasets The effectiveness of the model is demonstrated qualitatively and quantitatively on different human pose datasets.

The work presented in this chapter was initially presented in:

"PI-Net: Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation", Wen Guo, Enric Corona, Francesc Moreno-Noguer, and Xavier Alameda-Pineda, In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

The code of this work is released at `https://github.com/GUO-W/PI-Net`, and the project page of this work could be found at `https://team.inria.fr/robotlearn/pi-net-pose-interacting-network-for-multi-person-monocular-3d-pose-estimation/`.

4.1 INTRODUCTION

Monocular 3D multi-person human pose estimation is a challenging task that involves estimating the 3D joints of multiple individuals from a single RGB image. This problem has gained significant attention from both the research and industrial communities due to its potential applications in various fields such as entertainment, medical diagnosis, sports technology, physical therapy, etc. Previous research in multi-person human pose estimation usually treat individuals as independent instances, and estimate their poses separately using top-down methods within separate bounding boxes. However, these approaches disregards the contextual information such as the presence and interactions of other individuals in a same scene. This makes all these approaches agnostic about the context information and specifically about the presence of other people [138, 122, 139, 118, 84, 164, 117, 17, 40].

However, in a multi-person scene where people interact with each other, the pose and motion of every person are typically dependent and correlated with the body posture of the people they are interacting with. Contextual information has been demonstrated to be valuable in various computer vision tasks, including object detection [43, 130, 14], motion prediction [37], and affordance estimation [38]. However, to the best of our knowledge, the use of contextual information has not been well explored in the context of 3D human pose estimation. In this chapter, we investigate how these interdependencies can be leveraged to enhance the performance of off-the-shelf architectures for 3D human pose estimation.

Concretely, we propose a pose interacting network, PI-Net, which is fed with the 3D pose of a person of interest and an arbitrary number of body poses from other people in the scene, all of them computed with a context agnostic pose detector. These poses are potentially noisy, both in their absolute position in space and in the specific representation of the body posture. PI-Net is built using a recurrent network with a self-attention module that encodes the contextual information. Since it is unclear how to rank the contextual information, that is the pose of other persons, regarding the potential impact on the pose refinement pipeline, we make the very straightforward assumption that the potential of a person to refine the pose of the person of interest, is inversely proportional to the square of the distance between them.

We thoroughly evaluate our approach on the MuPoTS dataset [118], and using the initial detections of 3DMPPE [122], the current best performing approach on this dataset. PI-Net exhibits a consistent improvement of the pose estimates provided by 3DMPPE in all sequences of the dataset, becoming thus, the new state-of-the-art (see one example in Fig. 4.1). Interestingly, note that PI-Net can be used as a drop-in replacement for any other architecture that estimates 3D human pose. Additionally, the size of the network we propose is relatively small ($3.41M$ training parameters, while the baseline model has $36.25M$ parameters), enabling efficient training and introducing a marginal computational cost at the test. Testing on one Geforce1070, PI-net just cost 0.007s on refining one person while the baseline cost 0.038s for detecting one root-centered pose and also extra time on obtaining the bounding boxes and roots. Our method is lightweight and consistently improves the baseline.

4.2 RELATED WORK

Multi-person pose prediction from single RGB image 2D human pose estimation has been well explored under two kinds of approaches: the top-down approach and the bottom-up approach. Top-down approaches [171, 99, 148, 32] use a detector to get the bounding box of each person, and then use a single-person pose estimation network on each detected area. Bottom-ups [30, 29, 135, 126] approach takes a whole image as input which contains multi-person, detect all the locations of each joint (i.e. all the heads, all the knees, etc. in the images.) and then group the joints which belong to a same person together.

When coming to 3D, the problem becomes more difficult as the depth of each joint should also be predicted but not only the XY-location of the joints on the image, but the input image does not have depth signals. 3D pose estimation for single person either learn 3D poses directly from an input image [115, 98, 132, 136, 35], either estimates 2D poses first and than lift the 2D estimated poses to 3D with the learnt depth prediction [168, 20, 123]. And when studying 3D pose estimation in a multi-person scenario, similar to 2D multi-person pose estimation, the problem could also be solved in top-down [138, 139, 122, 84, 169] or bottom-up [118, 117, 176] approaches.

Contextual information in pose estimation Despite the fact that the above works estimate the body pose of an arbitrary number of people, each person is processed using an independent pipeline that does not take into account the interactions between the rest of the people or other contextual information. People never occur in isolation but are always related to the surrounding people or environment, thus some recent works begin to consider contextual information for pose estimation [175, 77, 92]. For example, Jiang *et al.* [77] defined a depth-ordering aware loss to take the order of the person involved in the occlusion, Li *et al.* [92] defined 3 different losses for three different levels of the human body to consider the depth order of different people and different joints. These works define different losses to consider the depth order of different people, but they do not really make use of the interaction relationship of different people. Fieraru *et al.* [47] proposed a person interaction dataset along with a framework based on contact detection over model surface regions, but this dataset is not released at the moment this work is published.

In this chapter, we propose a method that can be used in combination with the current state-of-the-art model [122] and boost its performance by looking at the whole group of humans and taking advantage of the signals of their surroundings. The proposed model is flexible and can be stacked after any 3D pose estimation model, independently of it being top-down or bottom-up.

4.3 PI-NET FOR MULTI-PERSON POSE ESTIMATION

Our goal is to exploit the interaction information between N people so as to improve the estimation of their pose. We assume the existence of an initial 3D pose estimate $\mathbf{P}_n \in \mathbb{R}^{J\times 3}$ of person $n = 1, \ldots, N$, where J is the number of estimated joints, e.g. obtained from 3DMPPE [122]. All the N poses are in absolute camera coordinates.

Formally, our goal is to improve the initial pose estimates, taking into account the pose of other people:

$$[\mathbf{Q}_1, \ldots, \mathbf{Q}_N] = \mathbf{\Pi}(\mathbf{P}_1, \ldots, \mathbf{P}_N), \tag{4.1}$$

where $\mathbf{Q}_n \in \mathbb{R}^{J\times 3}$ denotes the pose of person n improved with the information of the poses of the interactees.

While the idea is very intuitive, the research question is how to design PI-Net (i.e. $\mathbf{\Pi}$) so that it satisfies the following desirable criteria. Firstly, it shall work in environments with different numbers of people N, and not be fixed to a particular scenario. Secondly, the interaction information can be efficiently exploited and learned using publicly available datasets. Finally, it has to be generic enough to work with *any* 3D monocular multi-person pose estimator.

4.3.1 PIPELINE OF PI-NET

Naturally, the fact that the number of people N is unknown in advance, points us toward the use of recurrent neural networks. Such RNN should input the poses estimated by a generic pose estimator and embed the pose information into a representation learned specifically to take the cross-interactions into account. Without loss of generality, let us assume that the person-of-interest is $n = 1$, and hence the pose to refining is $\mathbf{P}_1$. We

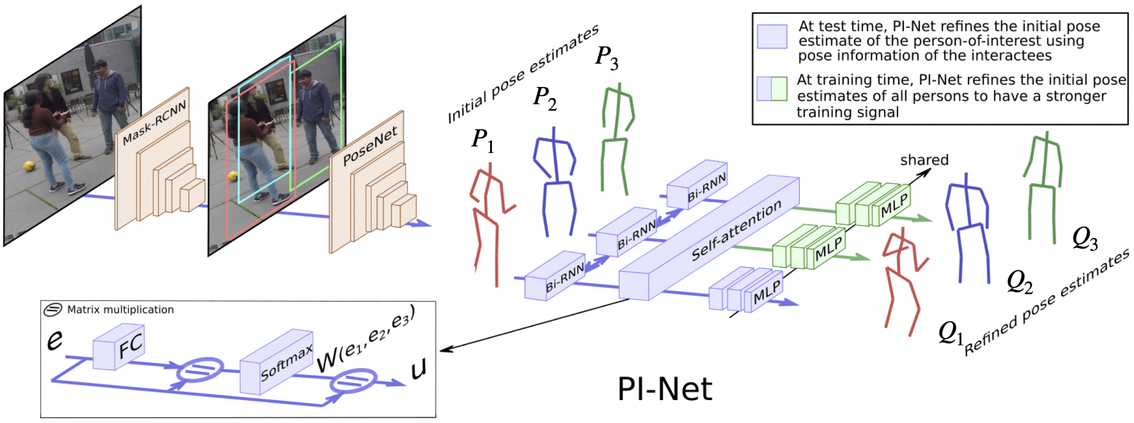

At test time, PI-Net refines the initial pose estimate of the person-of-interest using pose information of the interactees
At training time, PI-Net refines the initial pose estimates of all persons to have a stronger training signal
Mask-RCNN
PoseNet
Initial pose estimates
P_1
P_2
P_3
Bi-RNN
Bi-RNN
Bi-RNN
Self-attention
MLP
MLP
MLP
shared
Q_1
Q_2
Q_3
Refined pose estimates
Matrix multiplication
e
FC
Softmax
$W(e_1,e_2,e_3)$
u
PI-Net

Figure 4.2: **PI-Net Architecture**. Mask-RCNN [66] and PoseNet [122] are used to extract the initial pose estimates $\mathbf{P}_1, \dots, \mathbf{P}_N$. These estimates are fed into PI-Net, composed of three main blocks: Bi-RNN, Self-attention, and the shared fully-connected layers. The output of PI-Net refines the initial pose estimates by exploiting the pose of the interactions, yielding $\mathbf{Q}_1, \dots, \mathbf{Q}_N$.

consider using a bi-directional RNN, whose first input is $\mathbf{P}_1$, and then the rest of the initial poses are provided in a given order (see below). Our intuition for using a Bi-RNN is the following. During the forward pass, and since the first input is $\mathbf{P}_1$, the network can use the information in $\mathbf{P}_1$ to extract the features of the other poses that will best refine $\mathbf{P}_1$. In the backward pass, the network accumulates all this information back to $\mathbf{P}_1$, obtaining:

$$\mathbf{e}_1 = \text{Bi-RNN}(\mathbf{P}_1, \dots, \mathbf{P}_N). \tag{4.2}$$

The learned embedding $\mathbf{e}_1 \in \mathbb{R}^{N\times E}$is supposed to contain the crucial information from all other poses to refine the pose of the person-of-interest (1 in our example), but not only. Indeed, given that a priori we do not know which persons would be more helpful in refining the pose of interest, the computed embedding $\mathbf{e}_1$ could contain information that is not exploitable to refine the pose. In order to take this phenomenon into account, we soften the requirements of the Bi-RNN through the use of an attention mechanism as shown in Figure 4.2 (bottom-left zoom). Such attention mechanism aims to improve each embedding by combining information from the embeddings of other persons. To do so, we compute a matrix of attention weights:

$$\mathbf{W} \in \mathbb{R}^{N\times N}, \qquad \mathbf{W}_{nm} = \mathbf{e}_n^\top(\mathbf{A}_{\text{ATT}}\mathbf{e}_m + \mathbf{b}_{\text{ATT}})), \tag{4.3}$$

that is then normalized with a row-wise soft-max operation. $\mathbf{A}_{\text{ATT}}$ and $\mathbf{b}_{\text{ATT}}$ are attention parameters to be learned. The self-attention weights $\mathbf{W}$ encoding the residual interaction not captured by the Bi-RNN are used to update the embedding vector $\mathbf{u}_1 = \mathbf{W}\mathbf{e}_1$. Finally, the updated embedding is feed-forwarded through a few fully connected layers, obtaining the final refined pose $\mathbf{Q}_1$. While, at test time the self-attention and fully-connected layers are used only for the person of interest, at training time we found it is useful to apply these two operations to all poses and back-propagate the loss associated with everyone. This strategy eases the training. The overall pipeline depicting of PI-Net is shown in Figure 4.2.

4.3.2 INTERACTION ORDER

In the previous section, we assumed that the order in which the initial pose estimates $\mathbf{P}_n$ were presented to the Bi-RNN was given. Although there is no principled rule to define the ordering, there are some requirements. For a given person n, the sequence of poses presented to the network $\mathbf{P}_{\rho_n(1)}, \ldots, \mathbf{P}_{\rho_n(N)}$ has two constraints: (i) each pose is presented only once and (ii) the first pose is the one to be refined, i.e. $\rho_n(1) = n$. Intuitively, the order should represent the relevance: the more useful $\mathbf{P}_m$ is to refine $\mathbf{P}_n$, the closer $\mathbf{P}_m$ should be to $\mathbf{P}_n$ in the input sequence, i.e. the smaller $\rho_n(m)$ should be. Because finding the optimal permutation is a complex combinatorial optimization problem for which there is no ground-truth, we opt for assuming that the relevance is highly correlated to the physical proximity between interactees. Therefore, the closer person m is to person n, the smaller should $\rho_n(m)$ be. With this rule, we order the initial pose estimates to be fed to the Bi-RNN.

We also consider using Graph Convolutional Network [79] to model the interaction between different persons. Considering a pair of input persons, the node of the graph represents the coordinates of all the joints of these two people, and the adjacency matrix learned from the input represents the interaction between these joints. This strategy does not provide any performance increase, the results will be discussed in Section 4.4.4.

4.3.3 NETWORK ARCHITECTURE

In order to build and train our PI-Net, we first extract the initial poses using [122]. In the baseline, Mask-RCNN is used to detect the people present in the image. After that, the keypoint detector is applied to each image to detect the root-based poses and then project them into absolute camera coordinates. This keypoint detector is based on ResNet50 and 3 additional deconvolutional layers, following [150]. The set of keypoints for each person in camera coordinates $\mathbf{P}_n$, is therefore obtained. Note that this regressor gives all J person joints, despite partial occlusions, the corresponding occluded joints are hallucinated.

These initial pose estimates are then normalized with their mean and standard deviation, thus obtaining the input pose estimates of our PI-Net, $\{\mathbf{P}_1, \ldots, \mathbf{P}_N\}$. For each person n, we feed the PI-Net with the sequence of poses in the order appropriate for person n (see Section 4.3.2). The output $\mathbf{Q}_n$ of PI-Net is the refined pose for person n. PI-Net is trained with the L_1 loss between the refined poses and the ground-truth in 3D camera coordinates, added for all detected persons in the training image.

The Bi-RNN is implemented using three layers of gated recurrent units (GRU [34]). The self-attention layer provides a straightforward way to account for person-person interactions. After applying attention, the updated embedding goes through three fully connected layers to output the refined 3D pose in camera coordinates. These three fully connected layers are shared by all N poses. Consequently, the proposed PI-Net can be trained and evaluated using images with different numbers of people.

4.3.4 IMPLEMENTATION DETAILS

We use PoseNet of 3DMPPE [122] to generate our input 3D human pose. This model is trained on large-scale training data which includes H3.6M single-person 3D dataset [75], MPII[11] and COCO 2D dataset [101], MuCo multi-person 3D dataset [118], and extra synthetic data. PI-Net is trained on 33.4k composited MuCo data, which is contained in the training data of the baseline model. This ensures that the improvement of PI-Net compared with the baseline model is not caused by adding extra training data.

In terms of dimensions, 3DMPPE [122] outputs $J = 17$ joints in 3D, the hidden recurrent layers are of dimension 256, and the Bi-RNN outputs an embedding vector of dimension $E = 512$. We train our PI-net using Adam optimization and the *poly learning rate policy* [178], with an initial learning rate of 1e-5, a final learning rate of 1e-8, and power of 0.9, for 25 epochs. The batch size is set to 4.

When testing on an image with n instances, we test for n independent times, each time with a different ordering, and just retain the first person in each case.

4.4 EXPERIMENTS

We next describe the experiment section, which includes a description of the datasets, baselines, and evaluation metrics. We then provide a quantitative and qualitative evaluation and comparison to state-of-the-art approaches. We finalize this section with an exhaustive ablation study of the PI-Net architecture and hyperparameters.

4.4.1 DATASETS

MuCo-3DHP dataset and MuPoTS-3D dataset. The experiments discussed below are primarily based on two well-known datasets, introduced by Mehta et al. [118], that are widely used in the task of multi-person 3D human pose estimation. The MuCo-3DHP dataset is a multi-person 3D human pose dataset of multi-person composed by single-person data. As it is not real multi-person scenes, it is usually used for training but not testing. Our PI-Net is trained on $33.4k$ MuCo images, which contain $80.7k$ instances, without the need for additional data. The MuPoTS-3D test set consists of 8320 images from 20 real scenes, including 5 indoor and 15 outdoor scenes. In each scene, there are 2 or 3 people engaged in common activities such as talking, shaking hands, or doing sports. Each scene contains from 200 to 800 frames extracted from a video, and the dataset provides annotations in COCO format for both 2D image coordinates and 3D camera coordinates for each body joint.

COCO dataset. We also present qualitative results using the COCO dataset, which is a large-scale multi-person human pose dataset that provides only 2D ground truth labels. Despite this limitation, the dataset includes challenging scenes with a large number of people performing diverse actions, making it suitable for qualitative testing of models on complex in-the-wild cases. Specifically, we utilize examples from the COCO val2017 subset [66].

4.4.2 Baseline and Evaluation metrics

Our pipeline is capable of refining the poses estimated by any multi-person pose algorithm, independently of the strategy it uses. Given these initially estimated poses we refine them by leveraging the contextual information. In this chapter, we use the recent 3DMPPE [122] as a baseline and demonstrate both quantitative and qualitative improvements. Note that previous state-of-art works such as PandaNet [16] or SingleShot [118] do not provide codes either for training or testing and hence, we could not use them as backbones. The baseline [122] consists of 3 main steps. Firstly, 2D bounding boxes of humans are detected using Mask-RCNN [66]. For each detection, a deep network refines the coarse root 3D coordinates obtained from camera calibration parameters, and finally, a fully convolutional network [150] predicts root-relative 3D pose. Using the 3D root position, all poses can be represented in a common camera-coordinates reference.

We evaluate the performance of all methods by reporting the percentage of keypoints detected by the network that are within 150mm or less from the ground truth labels (3DPCK@150mm). This is the usual evaluation metric on the MuPOTS-3D test set[118, 138, 139, 117, 16, 122].

Notice that the 3DPCK metric depends greatly on the chosen threshold, for completeness, we also provide MPJPE and PA-MPJPE metrics to evaluate the performances. MPJPE indicates mean-per-joint-position error after root alignment with the ground truth [75], and PA-MPJPE denotes MPJPE after Procrustes Alignment[54]. Lower MPJPE and PA-MPJPE indicate better performance.

4.4.3 MAIN RESULTS

Quantitative results on MuPoTS-3D testset. We report the results of PI-Net on the MuPoTS-3D dataset in Table 4.1 and compare them to current state-of-the-art methods. Our results are obtained using the model depicted in Fig. 4.2, which uses a bidirectional 3-GRU recurrent layer, followed by a self-attention layer. We provide results after root alignment with the ground-truth poses, on the two strategies usually used on the MuPoTs datasets. In table4.1, the top-rows *Accuracy for all ground truths* evaluates all annotated persons, and the bottom rows *Accuracy only for matched ground truths* evaluates only predictions matched to annotations by their 2D projections with the 2D ground truths. We got improvements on both of the two strategies. PI-Net outperforms all previous models and improves the state-of-the-art by 1.3% 3DPCK@150mm on average. The improvement is consistent and shows a boost in performance for the majority of actions, setting a new state-of-the-art on the MuPoTS-3D dataset. Interestingly, we observe that the largest improvements are produced in those actions that require harmony and certain synchronization between people, such as practicing Taekwondo (S2) or playing a ball together (S14). We use ground-truth bounding box and roots to test the baseline, so the root-relative result is comparable with the absolute result here. To avoid redundancy, we only report root-relative results, which are widely reported in the previous works, for comparison with the state-of-the-art methods.

Table 4.2 shows the comparison of sequence-wise performance using MPJPE with root alignment and PA-MPJPE with further rigid alignment. Testing our model on the MuPoTS test dataset, we reduced the MPJPE error and PA-MPJPE error by 2.6mm and 4.3mm on average, respectively, in comparison with the baseline results [122]. Again, results are consistent across different tasks.

Table 4.3 shows a joint-wise comparison with state-of-art methods using 3DPCK@150mm after root alignment with ground truths. While we achieve similar performance with [122] in the head, neck, and hip, our method consistently outperforms the rest of the joints on arms and legs (shoulder, elbow, wrists, and knees). Arguably, the joints on the torso have little influence on the interaction between people, which comes mostly through the limbs, for example, hands and legs. Hence, it is reasonable that using the context information to refine 3D pose predictions gives the most significant boost in these joints.

Finally, it is worth pointing out that the results for all previous approaches reported in Tables 4.1, 4.2 and 4.3 are those of the respective papers. For 3DMPPE [122], however, we tested on ground-truth bounding boxes and roots to report these results.

Qualitative results on COCO. Figure 4.3 shows qualitative results on COCO dataset, for which 3D ground truths are not available. We also include (bottom-right) a failure case, caused by a misdetection of the baseline. This may be the major limitation of PI-Net, which is designed to refine poses, but so far, we have not integrated any module to deal with large deviations in the input poses.

4.4.4 ABLATION STUDY

Next, we provide a detailed analysis of the architectural design of PI-Net and discuss the interpretation of the predicted adjacency matrix obtained in the self-attention layers.

Effect of the Input Order. Table 4.4 shows the effect of using different strategies to establish the ordering of the detected people fed into the Bi-RNN layer.

We experimented with three different orders for processing the input people: (i) a random ordering, (ii) our approach where we select the person of interest followed by people in order of proximity, and (iii) the inverse approach where the person further away is processed first. To estimate the distance between people, we compute the distances between the root coordinates of the input people and the target person.

Table 4.1: Sequence-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, the fourth method and our model are tested with ground truth bounding boxes and roots. Higher value means better performance.

Sequence	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Accuracy for all ground truths										
LCR[138]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1
Singleshot[118]	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9
Xnect[117]	88.4	65.1	68.2	72.5	76.2	46.2	65.8	64.1	75.1	82.4
LCR++[139]	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5
PandaNet[16]	-	-	-	-	-	-	-	-	-	-
3DMPPE[122]	93.2	75.6	80.3	81.5	84.6	75.3	**84.5**	69.3	90.1	92.0
PI-Net (ours)	**93.5**	**77.4**	**82.0**	**82.9**	**87.2**	**75.9**	84.0	**71.5**	**90.2**	**92.2**
Accuracy only for matched ground truths										
LCR[138]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1
Singleshot[118]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9
LCR++[139]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5
Xnect[117]	88.4	70.4	68.3	73.6	82.4	46.4	66.1	83.4	75.1	82.4
3DMPPE[122]	**93.9**	83.0	80.3	81.5	85.4	75.3	84.5	77.2	90.1	92.0
PI-Net (ours)	**93.9**	**85.0**	**81.5**	**83.0**	**88.9**	**75.6**	**84.7**	**78.0**	**90.4**	**92.2**

Sequence	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	AVG
Accuracy for all ground truths											
LCR[138]	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Singleshot[118]	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Xnect[117]	74.1	72.4	64.4	58.8	73.7	80.4	84.3	67.2	74.3	67.8	70.4
LCR++[139]	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
PandaNet[16]	-	-	-	-	-	-	-	-	-	-	72.0
3DMPPE[122]	81.0	81.0	73.4	73.5	81.8	89.6	88.4	84.3	74.5	70.6	81.2
PI-Net (ours)	**82.5**	**82.9**	**74.7**	**75.7**	**83.6**	**91.4**	**90.6**	**86.0**	**74.9**	**71.1**	**82.5**
Accuracy only for matched ground truths											
LCR[138]	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Singleshot[118]	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	90.9	70.8
LCR++[139]	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Xnect[117]	76.5	73.0	72.4	73.8	74.0	83.6	84.3	73.9	85.7	90.6	75.8
3DMPPE[122]	81.0	81.0	74.3	76.0	81.8	89.6	88.4	84.3	75.5	76.2	82.6
PI-Net (ours)	**82.5**	**82.6**	**76.0**	**77.6**	**83.5**	**91.5**	**90.5**	**85.9**	**75.7**	**78.5**	**83.9**

Table 4.2: PA MPJPE (top) and MPJPE (bottom) comparisons of PI-net with the state-of-the-art method [122] used as our baseline on the MuPoTS dataset. The average value indicated the image-wise average. Ground truth bounding boxes and roots are used for testing. Lower value means better performance.

Sequence	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	AVG
PA MPJPE (mm)																					
3DMPPE [122]	67.7	102.6	82.7	82.5	79.8	**91.1**	70.8	110.1	72.8	63.5	88.6	79.6	105.1	110.5	77.5	72.2	73.3	86.8	91.9	120.0	88.4
PI-Net (ours)	**65.8**	**97.7**	**82.2**	**82.4**	**77.7**	91.6	**68.6**	**106.3**	**70.0**	**60.5**	**88.0**	**77.7**	**102.3**	**106.6**	**75.5**	**70.2**	**71.5**	**83.7**	**88.9**	**112.6**	**85.79**
MPJPE (mm)																					
3DMPPE [122]	90.9	159.3	121.8	113.5	107.8	**121.1**	113.8	138.2	99.7	98.4	119.6	115.4	143.7	151.7	111.7	101.8	105.6	115.8	140.7	187.7	126.0
PI-Net (ours)	**87.3**	**151.3**	**117.1**	**109.9**	**103.9**	**121.1**	**108.7**	**133.9**	**95.8**	**93.0**	**117.0**	**112.2**	**141.1**	**146.2**	**108.0**	**98.0**	**102.5**	**111.8**	**136.2**	**178.4**	**121.7**

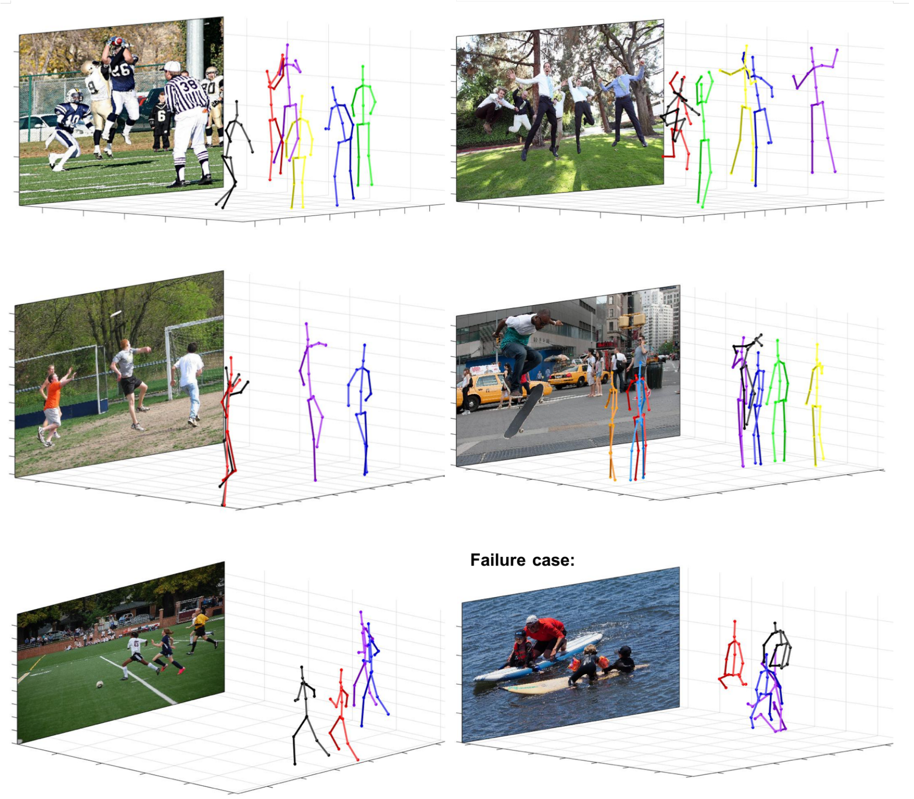


Figure 4.3: **Qualitative results on the COCO dataset**. For each pose, a darker color is used to represent the left side of the person. The bottom-right example corresponds to a failure case, as the ‘red’ and ‘black’ persons should be located in front of the scene, behind the ’blue’ and ’purple’ persons. This is caused by a misdetection on the root position of the input detected poses provided by the baseline network, while our network designed for refining the poses could not refine this kind of large deviation because this large deviation caused by the baseline network hinders our PI-net from learning the correct context information for correctly interpreting and refining the prediction.

Table 4.3: Joint-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, and the fourth method and our model are tested with ground truth bounding boxes and roots. All ground truths are used for evaluation. Higher value means better performance.

Method	Hd.	Nck.	Sho.	Elb.	Wri.	Hip	Kn.	Ank.	Avg
LCR[138]	49.4	67.4	57.1	51.4	41.3	84.6	56.3	36.3	53.8
single-shot[118]	62.1	81.2	77.9	57.7	47.2	**97.3**	66.3	47.6	66.0
3DMPPE[122]	**78.4**	**91.9**	83.1	79.7	67.0	93.9	84.3	**75.3**	81.2
PI-Net (ours)	78.3	91.8	**87.8**	**81.9**	**68.5**	94.2	**85.3**	74.8	**82.5**

Table 4.4: Comparison of different input orders. *Intuitive* is the one described in Section 4.3.2, from near to far. *Inverse* is the opposite. *Random* means in random order.

Order	PA MPJPE (mm)	MPJPE (mm)
Reverse	86.09	122.23
Random	85.87	121.88
Intuitive	**85.79**	**121.7**

Table 4.5: Importance of self-attention and bidirectionality (RNN). PI-Net uses a bidirectional RNN followed by a self-attention layer. We evaluate the impact of each of these choices: w/o Att. when removing attention, w/o Bi. considering standard RNN.

Method	PA MPJPE(mm)	MPJPE(mm)
PI-Net w/o Att., w/o Bi.	86.69	122.7
PI-Net w/o Bi.	86.42	123.10
PI-Net w/o Att.	85.92	122.01
PI-Net	**85.79**	**121.7**

Although the number of people in MuPoTS dataset images is relatively small, the processing order of each person's information has an impact on the model's performance. As shown in the table, the ordering we used provided the best performance, while the inverse order resulted in the worst performance. This finding highlights the importance of considering context information when processing input data.

Effect of self-attention and bidirectional RNN. In Table 4.5 we analyze the effect of

Table 4.6: Ablating the unit of the interaction network: None [122], Graph Convolutional Networks (GCN); LSTM and Gated Recursive Units (GRU), with (2,3,4) layers.

Interaction	PA MPJPE (mm)	MPJPE (mm)	# Par.
None [122]	88.36	126.0	133M
GCN	88.67	126.3	34M
2 LSTM	86.45	122.5	2.78M
3 LSTM	86.17	122.3	4.36M
4 LSTM	86.32	121.7	5.93M
2 GRU	86.27	122.2	2.23M
3 GRU	**85.79**	**121.7**	3.41M
4 GRU	85.96	122.2	4.59M

using the self-attention layer, which confirms that it helps to boost performance. We also study the attention weights predicted by the self-attention layer. These weights are, as expected, large at the diagonal, which corresponds to the self-interaction. The larger the distance between two people is, the smaller the weights tend to be. Table 4.5 also compares our approach which employs Bi-RNN with a standard (not bidirectional) RNN. The ablation of the recurrent unit is done later. Bi-RNN reduces 0.69mm of the MPJPE error and 0.77mm of the PA-MPJPE error, while the self-attention layers give an extra improvement of 0.31mm on MPJPE and 0.13mm on PA-MPJPE.

Interaction unit. In Table 4.6 we report results using alternative units to take the interaction into account. More precisely, Graph Convolution Network (GCN) and LSTM/GRU with different numbers of layers. For the experiment with GCN, we learned an adjacency matrix for every pair of persons and represented the interaction between them. We considered 4 GCN layers to obtain the refined poses. We also ablated the recurrent unit: GRU or LSTM [50]. Even though the MPJPE error of 4 LSTM layers is similar to that of 3 GRU layers, we considered the latter because it performs better after rigid alignment, and uses much fewer parameters which enables it to be trained more efficiently.

4.5 CONCLUSION

In this chapter, we introduce PI-Net, a pose-interacting network that refines initial 3D body poses predicted by any pose estimator by leveraging the mutual interaction that occurs in multi-person scenes. PI-Net utilizes three main building blocks: a bi-directional RNN, a self-attention module, and an MLP, to learn these interactions. The network is flexible, lightweight, and cost-efficient, and has the potential to improve other approaches for multi-person 3D human pose estimation, leading to a new state-of-the-art. This line of work focuses on improving perception results by exploring the interaction between people. One possible extension is to include other contextual information such as objects or structures to better understand human actions and explore different ways to interpret relationships in the scene. In this chapter, we verified the feasibility of considering interactions to improve the estimation of human poses. However, we identified two limitations of this work: (1) Due to data limitations, we were only able to refine and test on the MuCo, MuPoTS, and COCO datasets, which only contain loosely interacted scenes. It would be more informative to study human interaction in actions that involve strong interaction, such as acroyoga, dancing, or sports, but publicly accessible 3D data for such actions was not publicly accessible. (2) Analyzing human interaction based on only one frame can be confusing and misleading, as human poses are variable and can have multiple plausible possibilities based on one single observed instance. Future research should explore the potential of utilizing temporal information and studying sequences of human poses, which could provide more informative insights. These problems will be addressed in the next chapters.

CHAPTER 5

MULTI-PERSON EXTREME MOTION PREDICTION

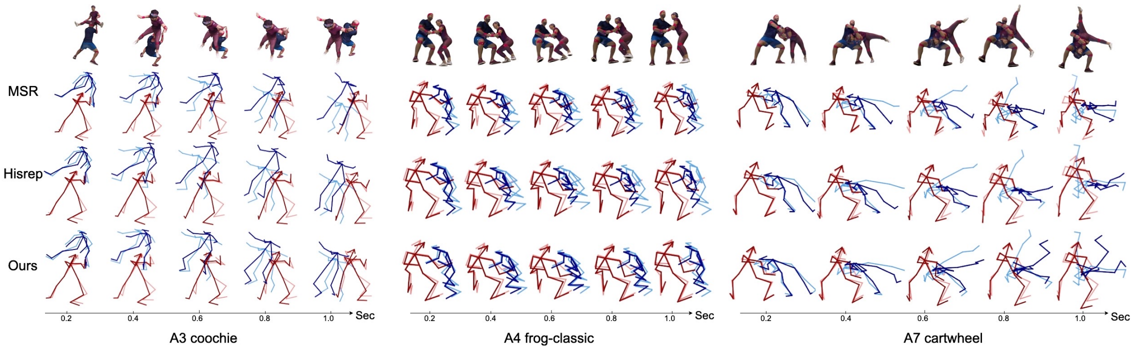


Figure 5.1: **Collaborative human motion prediction.** **1st row:** 3D sample meshes from our ExPI Dataset (just for visualization purposes). **2nd-4th rows:** Motion prediction results by MSR [42], Hisrep [111], and our method. Dark red/blue indicate prediction results, and light red/blue are the ground truth. By exploiting the interaction information, our approach of collaborative motion prediction achieves significantly better results than methods that independently predict the motion of each person.

Based on the limitations discussed at the end of the last chapter, we start to look into a sequence of human poses and begin to focus on the task of human motion prediction from now on. Human motion prediction aims to forecast future poses given a sequence of past 3D skeletons. While this problem has recently received increasing attention, it has mostly been tackled by single humans in isolation. In this chapter, we explore this problem when dealing with humans performing collaborative tasks, seeking to predict the future motion of two interacted persons given two sequences of their past skeletons. We investigate how to focus on the interaction of the two persons to improve the motion forecasting by a proposed baseline method with a Cross-interaction Attention (XIA) module that exploits the historical motion of two interacted persons to guide the prediction of their future movements. Besides, to solve the data lacking problem for this problem, we also proposed a new large dataset of highly interacted extreme dancing poses, called Extreme Pose Interaction Dataset(ExPI), along with a benchmark with three carefully selected train/test splits and two evaluation protocols. We verified the effectiveness of the proposed model on the ExPI dataset.

The work presented in this chapter was initially presented in: "Multi-person extreme motion prediction", Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-

Noguer, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

The code of this work is released at `https://github.com/GUO-W/MultiMotion`, and the project page of this work could be found at `https://team.inria.fr/robotlearn/multi-person-extreme-motion-prediction/`.

5.1 INTRODUCTION

The goal of human motion prediction is to predict future motions from previous observations. With the successful development of deep human pose estimation from single image [122, 61, 138, 139, 118, 84, 164, 117, 17, 40, 123], motion prediction begins to draw an increasing attention [36, 111, 4, 49, 56, 69, 76, 114, 15, 100, 52, 113, 140, 88]. Most existing works formulate motion prediction as a sequence-to-sequence task, where past observations of 3D skeleton data are used to forecast future skeleton movements. A common denominator of all these approaches is that they treat each pose sequence as an independent and isolated entity: the motion predicted for one person relies solely on her/his past motion. However, in real-world scenarios people interact with each other, and the motion of one person is typically dependent on or correlated with the motion of other people. Thus, we could potentially improve the performance of motion prediction by exploiting such human interaction.

Based on this intuition, in this chapter, we present a novel task: *collaborative motion prediction*, which aims to jointly predict the motion of two persons strongly involved in an interaction. To the best of our knowledge, previous publicly available datasets for 3D human motion prediction like 3DPW [164] and CMU-Mocap [55] that involve multiple persons only include weak human interactions, e.g., talking, shaking hands, etc. Here we move a step further and analyze situations where the motion of one person is highly correlated to the other person, which is often seen in team sports or collaborative assembly tasks in factories.

With the goal to foster research on this new task, we collected the ExPI (Extreme

Pose Interaction) dataset, a large dataset of professional dancers performing Lindy Hop aerial steps.[1] To perform these actions, the two dancers perform different movements that require a high level of synchronization. These actions are composed of extreme poses and require strict and close cooperation between the two persons, which is highly suitable for the study of human interactions. Some examples of this highly interacted dataset are shown in Figure 5.2. Our dataset contains 115 sequences of 2 professional couples performing 16 different actions. It is recorded in a multiview motion capture studio, and the 3D poses and 3D shapes of the two persons are annotated for all 30K frames. We have carefully created train/test splits and proposed two different extensions of the pose evaluation metrics for the collaborative motion prediction task.

To model such strong human-to-human interactions, we introduce a novel Cross-Interaction Attention (XIA) module, which is based upon a standard multi-head attention [160] and exploits historical motion data of the two persons simultaneously. For a pair of persons engaging in the same activity, XIA module extracts the spatial-temporal motion information from both persons and uses them to guide the prediction of each other.

We exhaustively evaluate our approach and compare it with state-of-the-art methods designed for single human motion prediction. Note that in our dataset of dancing actions, movements are performed at high speed. The long-term predictions are very challenging in this case. Nevertheless, the results demonstrate that our approach consistently outperforms these methods by a large margin, with $10 \sim 40\%$ accuracy improvement for short (≤ 500 ms) and $5 \sim 30\%$ accuracy improvement for long term prediction (500 ms ~ 1000 ms).

Our key contributions can be summarized as follows:

- We introduce the task of collaborative motion prediction, to focus on the estimation of future poses of people in highly interactive setups.
- We collect and make publicly available ExPI, a large dataset of highly interacted extreme dancing poses, annotated with 3D joint locations and body shapes. We also

---

[1] The Lindy Hop is an African-American couple dance born in the 1930s in Harlem, New York, see [121].

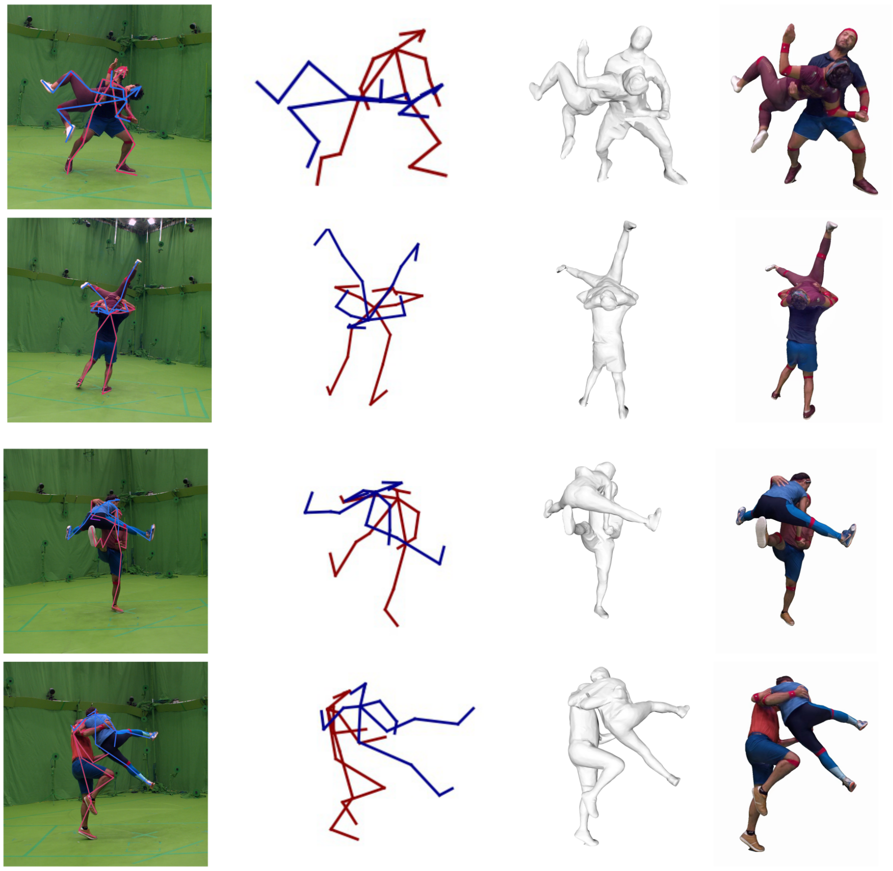


Figure 5.2: Some samples of the ExPI dataset: RGB image with projected 2D skeletons, 3D pose, mesh, and textured mesh.

define the benchmark with carefully selected train/test splits and evaluation protocols.

- We propose a method with a novel cross-interaction attention (XIA) module that exploits the historical motion of two interacted persons to predict their future movements. Our model can be used as a baseline method for collaborative motion prediction.

5.2 RELATED WORK

3D Human Motion Prediction Due to the inherent sequential structure of human motion, 3D human motion prediction has been mostly addressed with recurrent models. For instance, Fragkiadaki *et al*. [49] propose an encoder-decoder framework to embed human poses and an LSTM to update the latent space and predict future motion. Jain *et al.* [76] split the human body into sub-parts and forward them via structural RNNs. Martinez *et al.* [114] introduces a residual connection to model the velocities instead of the poses themselves. Interestingly, they also show that a model trained with diverse action data performs better than those trained with single actions. However, although RNNs achieve great success in motion prediction, they suffer from containing the entire history with a fixed-size hidden state and tend to converge to a static pose. Some works alleviate this problem by using RNN variants [103, 33], sliding windows [24, 25], convolutional models [70, 69, 89] or adversarial training [56].

Since the human body is a non-rigid and structured data, directly encoding the whole body into a compact latent embedding will neglect the spatial connectivity of human joints. To this end, Mao *et al.* [113] introduces a feed-forward graph convolutional network (GCN) [79, 161] with the learnable adjacent matrix. This approach was later boosted with self-attention on an entire piece of historical information [111] or a selection of them [91]. Recently, GCN-based methods are further developed by leveraging multi-scale supervision [42], space-time-separable graph [145], and contextual information [1, 2]. In terms of GCN design, Cui *et al.* [39] argues that training the adjacent matrix from scratch ignores the natural connections of human joints, and proposes to use a semi-constrained adjacent matrix. Li *et al.* [95] combines a graph scattering network with a hand-crafted adjacent matrix. Other works also exploit the use of transformers [160] to replace GCN in human motion prediction [26, 4].

Considering that human actions are essentially stochastic in the future, some works leverage on generative models (e.g. VAEs and GANs) [172, 180, 8, 174, 7, 27, 112, 133]. Nevertheless, although these models can generate diverse future motions, their prediction accuracy still needs to be further improved when compared to deterministic models.

Contextual Information in Human Interaction Context information has been proved to be useful in pose estimation[92, 65, 175, 167, 77, 61] and trajectory prediction [6, 46, 162, 74, 83, 90]. In motion prediction, human-object interaction [36] scene context *et al.*[28] have been taken into consideration. Besides, social context-aware motion prediction has also been studied [1, 2] but only in weak interacted scenes. In any event, none of these papers explores the situation we contemplate in this chapter, in which humans do perform highly interactive actions.

Datasets Using deep learning methods to study 3D human pose tasks relies on high-quality datasets. Most previous 3D human datasets are single person [75, 109, 143] or made of pseudo 3D poses [164, 119]. Other datasets which contain lab-based 3D data usually do not have close interactions [142, 102, 119, 55]. Recently, some works have started to focus on the importance of context information and propose datasets to model the interaction of synthetic persons with scenes [28]. Furthermore, Fieraru *et al.* [47] created a dataset of human interaction with a contact-detection-aware framework, but this dataset just contains several daily scenarios with mild human interactions and it is not released yet at the time of the publishing of our work concerned in this Chapter. Thus, we believe the ExPI dataset we present here, where the actions of people are highly correlated, fills an empty space in the current datasets of human 3D pose/motion.

5.3 PROBLEM FORMULATION

As discussed in the introduction, the task of single-person human motion prediction is well established. It is defined as learning a mapping $\mathcal{M} : P_{t_{\text{I}}:t-1} \to P_{t:t_{\text{E}}}$ to estimate the future movements $P_{t:t_{\text{E}}}$ from the previous observation $P_{t_{\text{I}}:t-1}$, where t_{I} (t_{E}) denotes the initial (ending) frame of a sequence, and P_t denotes the pose at time t.

In this work, we extend the problem formulation to collaborative motion prediction of two interacted persons. While our formulation is general and could work for any kind of interaction, for the sake of consistency throughout the paper, we will denote by ℓ and f variables corresponding to the leader and the follower respectively (see Section 5.4 on the dataset description). Therefore, the collaborative motion prediction task is defined as learning a mapping:

$$\mathcal{M}_{\text{C}} : P^{\ell}_{t_{\text{I}}:t-1}, P^{f}_{t_{\text{I}}:t-1} \to P^{\ell}_{t:t_{\text{E}}}, P^{f}_{t:t_{\text{E}}}. \tag{5.1}$$

Since the two persons are involved in the same interaction, we believe it is possible to better predict the motion of a person by exploiting the pose information of her/his interacted partner. From now on, we will use $P_t^c = [P_t^l, P_t^f]$ to denote the joint pose of the couple (two actors) at time t, and P_t to denote either of them.

In the following parts of the paper, we will provide an experimental framework for the collaborative motion prediction task, consisting of a dataset and evaluation metrics, to foster research in this direction. And we will also introduce our proposed method for this task.

5.4 THE EXTREME POSE INTERACTION DATASET

We present the Extreme Pose Interaction (ExPI) Dataset, a new person interaction dataset of Lindy Hop dancing actions. In Lindy Hop, the two dancers are called *leader* and *follower*.[2] We recorded 2 couples of dancers in a multi-camera setup equipped with a motion-capture system. In this section we will first describe the recording procedure and data cleaning; then we will give a comprehensive analysis of our dataset components; finally, we will analyze the dataset with defined matrices to show the diversity and extremeness of our collected data.

5.4.1 DATASET COLLECTION AND POST-PROCESSING.

Data collection The data were collected in a multi-camera platform equipped with 68 synchronized and calibrated color cameras and a motion capture system with 20 mocap cameras.[3] The data collection is marker-based, which is to say, each actor is dressed with 18 different markers on the corresponding joints for the system to track and record.
Our data collection strategy went through an Ethics Review Board, and the recordings were authorized, together with the associated Consent Form. Our data does not contain any personally identifiable information beyond the images themselves. The data will be

---

[2] This is the standard gender-neutral terminology for Lindy-Hop.
[3] Kinovis `https://kinovis.inria.fr/`

shared respecting all national and international regulations, as authorized by COERLE, the Ethics Review Board at INRIA.

Data Post-processing When collecting the motion capture data, some points are missed by the system due to occlusions or tracking losses, which is a common phenomenon in lab-based interacted Mocap datasets [47]. This problem becomes even worse when dealing with extreme poses. To overcome this issue and ensure the quality of the data, we designed and implemented a 3D hand-labeling toolbox, and we spent months manually labeling the missing points.

To label the missing joints, for each missed value we choose two orthogonal views among the several viewpoints and label the missed keypoints by hand on these two frames to get two image coordinates. We then use the camera calibration to back project these two image coordinates into the 3D world coordinate, obtaining two straight lines. Ideally, the intersection of these two lines is the world coordinate of this missing point. Since these two lines do not always intersect, we find the nearest point, in the least-squares sense, to these two lines to approximate the intersection. Figure 5.3 shows an illustration of the idea of this calibration. In this procedure, we did not use the distortion parameters, since we observed that the distortion error is negligible on the views we choose for labeling. The intersection is projected into 3D and various 2D images to confirm the quality of the approximation by visual inspection. Figure 5.4 shows an example of before and after labeling the missing joints.

5.4.2 DATASET STRUCTURE

Components. 16 different actions are performed in ExPI dataset, some by the 2 couples of dancers, some by only one of the couples. Each action was repeated five times to account for variability. More precisely, for each recorded sequence, ExPI provides:

- Multi-view videos at 25FPS from all the cameras in the recording setup;
- Mocap data (3D position of 18 joints for each person) at 25FPS synchronized with

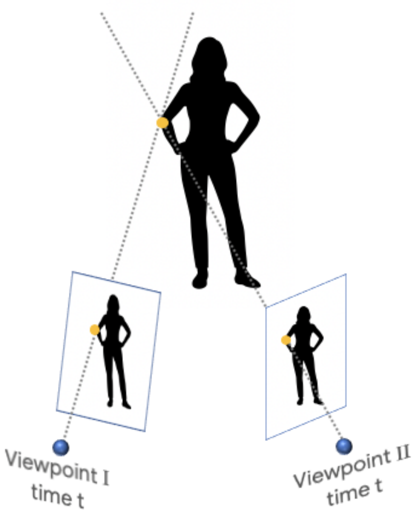


Figure 5.3: An illustration of labeling the missing joints[4].

the videos.;

- camera calibration information;
- 3D shapes as textured meshes for each frame.

Overall, ExPI contains 115 sequences, each one depicting an execution of one of the actions. It has in total 30k visual frames for each of the 68 viewpoints, and 60k 3D instances annotated.

Action names and joint order Table 5.1 shows the names of the 16 actions performed by the couples of actors in ExPI. In the video of the supplementary material, we include example videos for each of the 16 actions.
In the ExPI dataset, the pose of each person is annotated with 18 keypoints, so we have 36 keypoints for both actors. The order of the keypoints is as follows, where “F” and “L” denote the Follower and the Leader respectively, and “f”, “l” and “r” denote “forward”, “left” and “right”:

---

[4] This figure is coming from `https://ai.googleblog.com/2019/05/moving-camera-moving-people-deep.html`

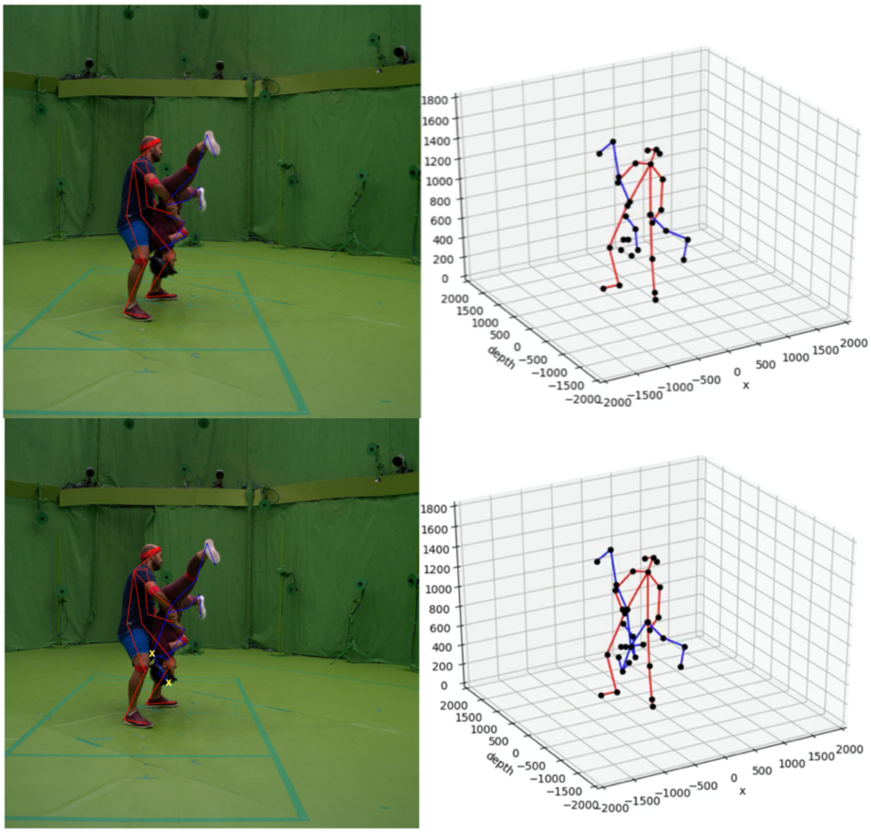

1800
1600
1400
1200
1000
800
600
400
200
0
depth
2000
1500
1000
500
0
−500
−1000
−1500
−2000
x

Figure 5.4: Data-cleaning. **Top:**Data before cleaning. The two joints 'F-back' and 'F-fhead' are missed. **Bottom:** Data after cleaning. The yellow marks indicate the two relabeled joints.

(0) 'L-fhead'
(1) 'L-lhead'
(2) 'L-rhead'
(3) 'L-back'
(4) 'L-lshoulder'
(5) 'L-rshoulder'
(6) 'L-lelbow'
(7) 'L-relbow'
(8) 'L-lwrist'
(9) 'L-rwrist'
(10) 'L-lhip'
(11) 'L-rhip'
(12) 'L-lknee'
(13) 'L-rknee'
(14) 'L-lheel'
(15) 'L-rheel'
(16) 'L-ltoes'
(17) 'L-rtoes'
(18) 'F-fhead'
(19) 'F-lhead'
(20) 'F-rhead'
(21) 'F-back'
(22) 'F-lshoulder'
(23) 'F-rshoulder'
(24) 'F-lelbow'
(25) 'F-relbow'
(26) 'F-lwrist'
(27) 'F-rwrist'
(28) 'F-lhip'
(29) 'F-rhip'
(30) 'F-lknee'
(31) 'F-rknee'
(32) 'F-lheel'
(33) 'F-rheel'
(34) 'F-ltoes'
(35) 'F-rtoes'

Table 5.1: Composition of the ExPI Dataset. The seven first actions are performed by both couples. Six more actions are performed by Couple 1, while three others by Couple 2.

Action	Name	Couple 1	Couple 2
A_1	A-frame	✓	✓
A_2	Around the back	✓	✓
A_3	Coochie	✓	✓
A_4	Frog classic	✓	✓
A_5	Noser	✓	✓
A_6	Toss out	✓	✓
A_7	Cartwheel	✓	✓
A_8	Back flip	✓	
A_9	Big ben	✓	
A_{10}	Chandelle	✓	
A_{11}	Check the change	✓	
A_{12}	Frog-turn	✓	
A_{13}	Twisted toss	✓	
A_{14}	Crunch-toast		✓
A_{15}	Frog-kick		✓
A_{16}	Ninja-kick		✓

Comparison with other datasets Table 5.2 compares our dataset with several other publicly available 3D human datasets that are widely used in recent work [114, 113, 111, 42]. From this table, we can see that our dataset is eminently suitable for the task of multi-person extreme motion prediction, and it is also able to be used in human pose estimation in rare conditions and challenging human shape estimation.

5.4.3 DATA ANALYSIS

Diversity. Similar to Ionescu *et al.* [75], we analyze the diversity of our dataset by checking how many *distinct* poses have been obtained. We consider two poses to be *distinct*, if at least one of the J joints for one pose P^c_m is different from the corresponding joint of the other pose P^c_n, beyond a certain tolerance τ (mm):

$$\max_{j\in[1,J]} \|P^c_{m,j} - P^c_{n,j}\| > \tau, \tag{5.2}$$

Table 5.2: Comparison of ExPI with other publicly available datasets commonly used for human motion prediction.

Dataset	AMASS[109]	H3.6m[75]	3DPW[163]	MuPoTS[119]	ExPI
3D joints	✓	✓	✓	✓	✓
Video	✓	✓	✓	✓	✓
Shape	✓	✓	✓		✓
Multi-person			✓	✓	✓
Extreme poses	✓				✓
Multi-view					✓

where $m, n \in \mathcal{D}$ denote any two poses in the dataset $\mathcal{D}$. Then we define *diversity* of the dataset as the percentage of *distinct* poses among all the poses. According to Ionescu *et al.* [75], the diversity of H3.6Mis 24% and 12% when setting the tolerance τ to 50 mm and 100 mm, respectively. While the diversities of ExPI for the same threshold values are 52% and 23%, which are much more diverse.

Extremeness. To measure the extremeness of a pose sequence, we first compute the standard deviation (*std*) over time for each dimension of the $xyz-$coordinate for every joint. Then, the extremeness of the joint j is defined as its maximum per-coordinate standard deviation: $\varepsilon_j = \max\{\sigma_j^x, \sigma_j^y, \sigma_j^z\}$. Finally, the extremeness of action is evaluated by computing the percentage of joint extremeness values ε_n within various intervals $[\varepsilon_{\min}, \varepsilon_{\max}]$. Figure 5.5 and Figure 5.6 reports the extremeness of ExPI dataset compared to H3.6M in two different ways: (i) a per-action plot reporting extremeness on various color-coded intervals (Figure 5.5); (ii) computing the percentage of joints more extreme than a certain *std* value(Figure 5.6). From both plots, it is clear that the ExPI dataset is significantly more extreme than the H3.6M dataset.

Quantitative tests of multiple tasks on ExPI dataset As presented in Section 5.4.2, ExPI provides various information containing multi-view RGB videos, Mocap 3D labels, textured mesh etc. Thus it would be proper for the use of multiple human understanding tasks such as pose estimation, motion estimation, detection and segmentation, etc. From

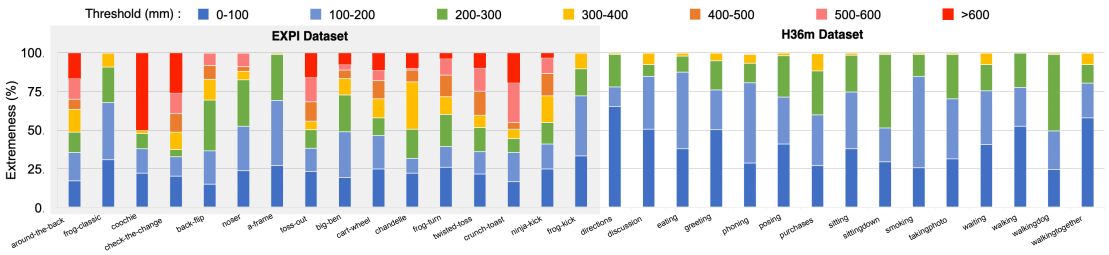


Figure 5.5: Per-action extremeness comparison of Human3.6M dataset and ExPI dataset. This figure shows the percentage of joints whose *std* is among a certain threshold (in different colors), for different actions. Actions with more red colors are more extreme.

the above analysis, we could see that ExPI dataset is extreme and diverse and, thus should be challenging for different tasks. To verify this, we use the official repos of several different state-of-the-art methods for multiple tasks for some quick quantitative tests on our collected data, see Figure 5.7. Note that here we use the pre-trained models provided by these works, but do not fine-turn on our training data. We could see that these methods perform badly when the poses are extreme, when the two people closely interact, or when occlusion is severe. One reason is that these methods are trained on publicly released datasets, though the amount of data is huge, there are rare samples for extremely interacted poses like our dataset. Also, these methods are not designed for such kind of highly interactive conditions, resulting in bad performance. Both these two reasons could support the conclusion that the ExPI dataset could serve as a meaningful and challenging dataset in this domain.

5.5 METHOD

We introduce our approach for collaborative motion prediction, aiming to set the first performance baseline to help future developments.

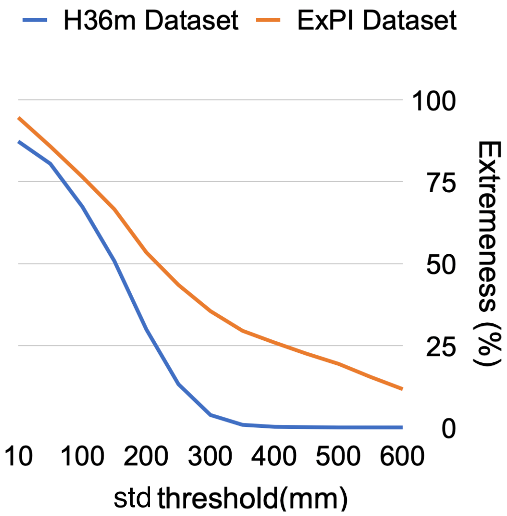
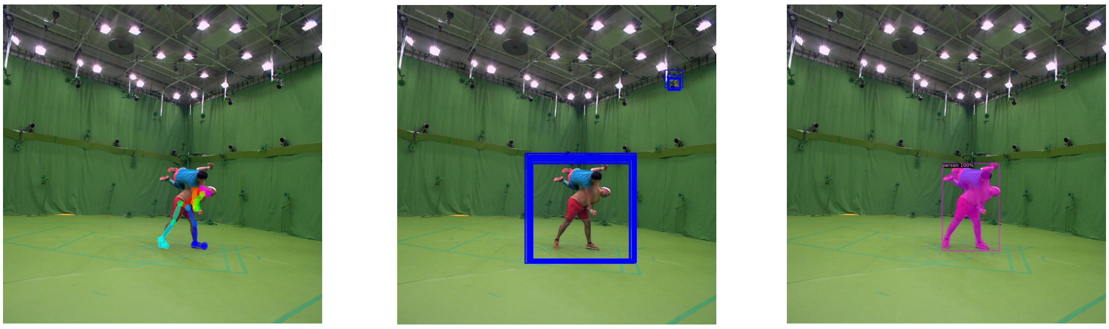


Figure 5.6: Average extremeness comparison of Human3.6M dataset and ExPI dataset. This figure shows the percentage of joints whose *std* is beyond a certain threshold.

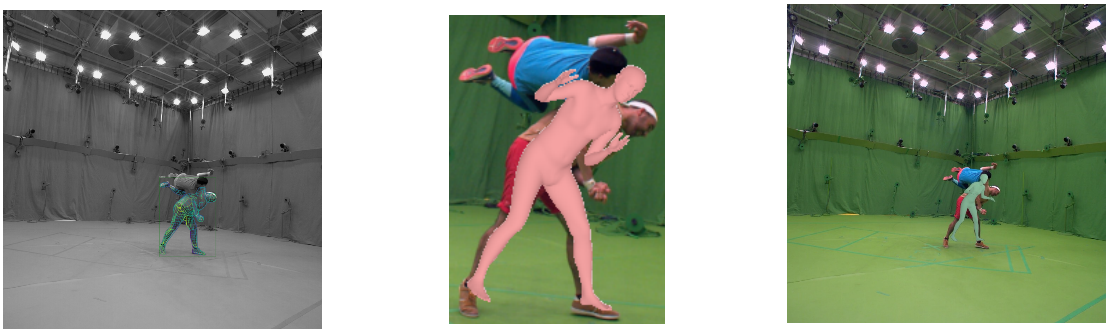
5.5.1 PIPELINE

The idea of our method is to learn two person-specific motion prediction mappings and to propose a strategy to share information between these two mappings. The possibility to include information from the other person involved in the interaction should push the network to learn a better representation for motion prediction. The overall pipeline is described in Figure 5.8.

For the two single-person motion prediction mappings, we draw inspiration from [111], using an attention model for leaning temporal attention w.r.t. the past motions, and a predictor based on Graph Convolutional Network (GCN) [79] to model the spatial attention among joints using an adjacency matrix. The temporal attention model aims to find the most relative sub-sequence in the past by measuring the similarity between the last observed sub-sequence and a set of past sub-sequences. In this attention model, the query Q is learned by MLP from the last observation $P_{t-1-M:t-1}$ (blue dashed rectangle in Figure 5.8, length M). The keys K_i are learned by MLP from the starting chunk of sub-sequences $P_{t_i:t_i+M}$(red dashed rectangles in Figure 5.8, length M). And the values



(a) Openpose 2D (b) Mask-rcnn detection (c) Mask-rcnn segmentation



(d) dense pose (e) SPIN (f) VIBE(maskrcnn based)

Figure 5.7: Quantitative tests of multiple tasks on ExPI dataset: (a) 2D pose estimation by Openpose [29]; (b) Person detection by Mask-rcnn [66]; (c) Instance segmentation by Mask-rcnn [66]; (d) Pose estimation by Dense pose [57]; (e) 3D pose reconstruction by SPIN [81]; (f) Human shape estimation by VIBE [80]

V_i consist of DCT representations built from the sub-sequences $P_{t_i:t_i+M+T}$ (black dashed rectangles in Figure 5.8, length $M+T$), where t_i with $i \in \{1, \ldots, N\}$ indicates the start frame of each past sub-sequence.

Training such a strategy separately for each actor does not account for any interaction between the two dancing partners. To deal with this, we design a cross-interaction attention (XIA) module based on multi-head attention, to introduce guidance from the interacted person. In the next section, we introduce this XIA module.

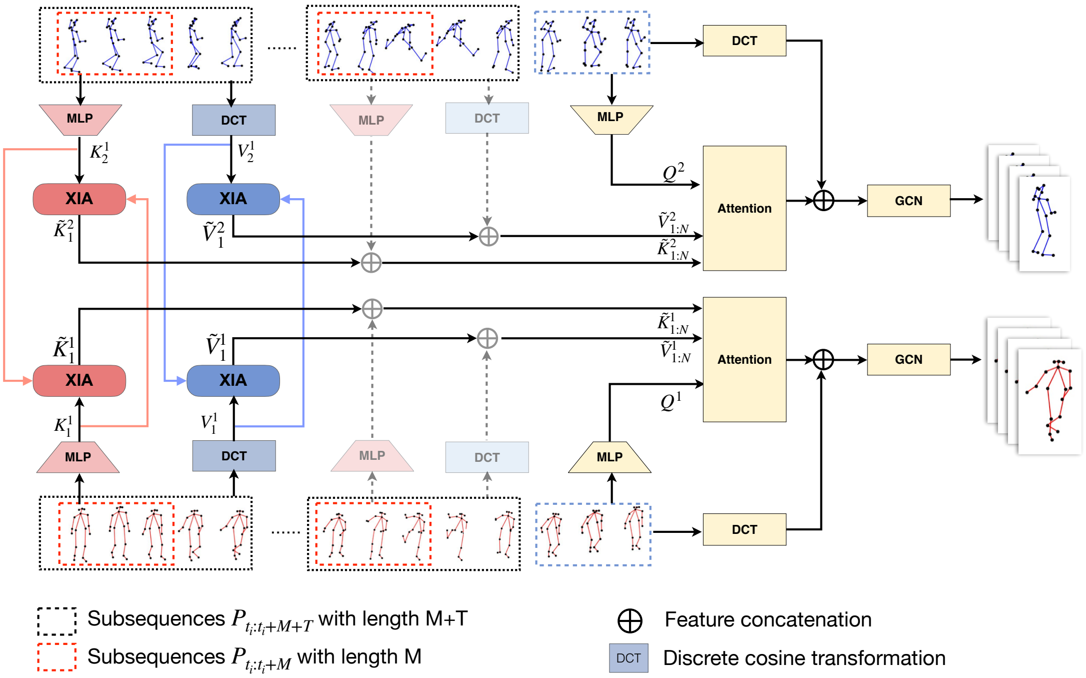

MLP
DCT
MLP
DCT
MLP
DCT
K_2^1
V_2^1
XIA
XIA
$\tilde{K}_1^2$
$\tilde{V}_1^2$
Q^2
Attention
$\tilde{V}^2_{1:N}$
$\tilde{K}^2_{1:N}$
GCN
$\tilde{K}^1_{1:N}$
$\tilde{V}^1_{1:N}$
Attention
GCN
Q^1
$\tilde{K}_1^1$
$\tilde{V}_1^1$
XIA
XIA
K_1^1
V_1^1
MLP
DCT
MLP
DCT
MLP
DCT
Subsequences $P_{t_i:t_i+M+T}$ with length M+T
Subsequences $P_{t_i:t_i+M}$ with length M
Feature concatenation
DCT Discrete cosine transformation

Figure 5.8: Computing flow of the proposed method. Two parallel pipelines – for the leader and the follower – are implemented. The key-value pairs are refined by XIA modules(we just visualize XIA modules for the first sub-sequences, while it is the same for the following sub-sequences).

5.5.2 CROSS-INTERACTION ATTENTION (XIA)

XIA aims to share motion information between the two predictors. In particular, we denote the query and the key-value pairs for one person by Q and $\{K_i, V_i\}_{i=1}^N$ respectively, and use the superscript f and ℓ to indicate the two persons, follower and leader. We naturally cast the collaborative human motion prediction task into learning how to jointly exploit the information in (K_i, V_i) when querying with Q to predict the motion of each person.

Our intuition is that the pose information (key-value pairs) of one person can be used to transform the pose information of the other person for better motion prediction. We implement this intuition with the help of the proposed *cross-interaction attention module*.

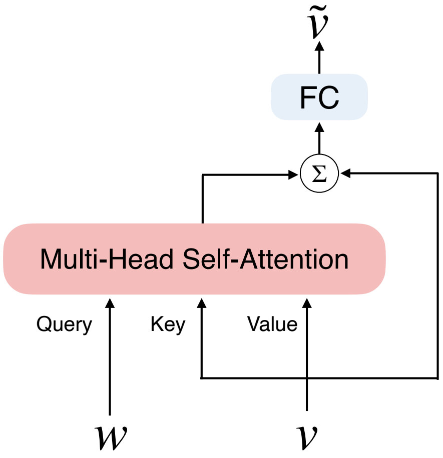


Figure 5.9: Cross-interaction attention (XIA) module. In order to refine w with the help of the corresponding interaction information $w_{int.}$, the multi-head attention is queried by $w_{int.}$ and take w as key and value.

Such a module takes as input w and the corresponding vector from the interacted pose $w_{int.}$, and uses multi-head self attention to get the refined vector $\tilde{w}$ (see Figure 5.9):

$$\tilde{w} = \text{XIA}(w_{int.}, w) = \text{FC}(\text{MHA}(w_{int.}, w, w) + w), \tag{5.3}$$

where $\text{MHA}(q, k, v)$ stands for multi-head attention with query q, key k and value v, and FC indicates fully connected layers. We use different XIA modules to update keys and values mentioned in Section 5.5.1: in our implementation, XIA modules for keys have 8 attention heads, and XIA for values has a single attention head. Moreover, we add a skip connection for the MHA module followed by 2 FC layers. XIA modules for leader/follower do not share weights.

The proposed XIA module is integrated at several stages of the computing flow as shown in Figure 5.8. More precisely, we refine all keys:

$$\tilde{K}_i^\ell = \text{XIA}(K_i^\ell, K_i^f), \qquad \tilde{K}_i^f = \text{XIA}(K_i^f, K_i^\ell), \tag{5.4}$$

and analogously for the values.

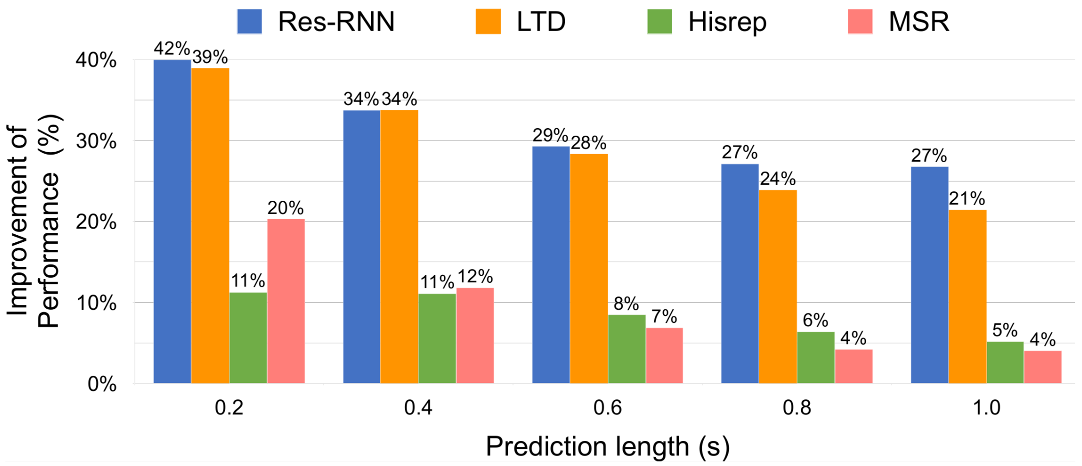
XIA could be potentially generalized to any number of participants by considering either several XIA modules and fusing their outcome or performing the fusion at the input of XIA module.

5.5.3 POSE NORMALIZATION

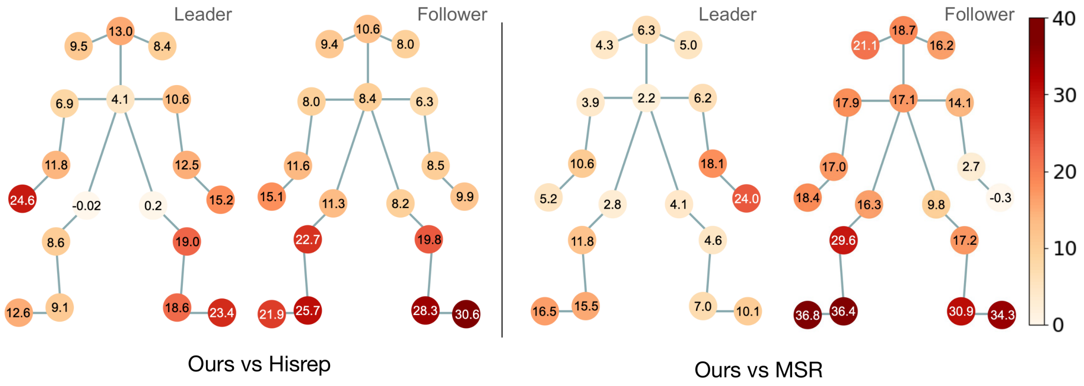
Raw poses of ExPI are represented in the world coordinate. Similar to single-person motion prediction, we normalize the data by removing the global displacement of the poses based on a selected root joint. While our task aims at predicting not only the distinct poses but also the relative position of the two persons, so we have to normalize by the same person to keep the information of their related position. We could either normalize by leader/follower or choose to normalize by the leader for better visualization. Specifically, for each frame, we take the root joint (middle of the two hips) of the leader as coordinate origin, use the root point and left hip of the leader to define x-axis, and use the neck of the leader to determine the XOZ plane. We normalize all the joints of both persons to this coordinate, then the pose errors can be computed directly in this coordinate. More precisely, we represent the raw poses in the world coordinate as $P_w \in \{P_w^\ell, P_w^f\}$, and $T_{P_{w,t}^\ell}$ is the rigid transformation aligning the two actors to the leader's coordinate system. The normalized coordinates are thus $P_t^\ell = T_{P_{w,t}^\ell} P_{w,t}^\ell$, and $P_t^f = T_{P_{w,t}^\ell} P_{w,t}^f$. In the following, P shall always represent the normalized pose unless specified otherwise.

5.6 EXPERIMENTAL EVALUATION

This section describes the experimental protocol on ExPI, and discusses the results of our proposed method.



(a) Percentages of improvement of our method comparing with different state-of-the-art methods


Leader
Follower
Leader
Follower
13.0
9.5
8.4
6.9
4.1
10.6
11.8
12.5
24.6
-0.02
0.2
15.2
8.6
19.0
12.6
9.1
18.6
23.4
10.6
9.4
8.0
8.0
8.4
6.3
11.6
8.5
15.1
11.3
8.2
9.9
22.7
19.8
21.9
25.7
28.3
30.6
6.3
4.3
5.0
3.9
2.2
6.2
10.6
18.1
5.2
2.8
4.1
24.0
11.8
4.6
16.5
15.5
7.0
10.1
18.7
21.1
16.2
17.9
17.1
14.1
17.0
2.7
18.4
16.3
9.8
-0.3
29.6
17.2
36.8
36.4
30.9
34.3
40
30
20
10
0
Ours vs Hisrep
Ours vs MSR

(b) Joint-wise JME improvement (mm)

Figure 5.10: (a): Percentages of improvement of our method comparing with different state-of-the-art methods, measured by average JME error on the common action split, at different forecast times. Lower value means closer performance with our model. Our method surpasses these methods up to $10 \sim 40\%$ in the short term, and $5 \sim 30\%$ in the long term. (b): Joint-wise JME improvement(mm) of our method over Hisrep [111] and MSR [95]. Darker color means larger improvement.

5.6.1 SPLITTING THE EXPI DATASET

As described in Sect. 5.4.2, we record 16 actions in ExPI dataset. Seven of them are common actions (A_1 to A_7) which are performed by both of the 2 couples: we denote them as $\mathcal{A}_c^1$ performed by couple 1 and $\mathcal{A}_c^2$ by couple 2. The other actions are couple-specific, which are performed only by one couple: we denote the actions performed by couple 1 (A_8 to A_{13}) as $\mathcal{A}_u^1$, and actions by couple 2 (A_{14} to A_{16}) as $\mathcal{A}_u^2$. With these notations, we propose three data splits.

Common action split. Similar to [75], we consider the common actions performed by different couples of actors as train and test data. More precisely, $\mathcal{A}_c^2$ is the train dataset and $\mathcal{A}_c^1$ is the test dataset. Thus, train and test data contain the same actions but are performed by different people.

Single action split. Similar to [49, 76], we train 7 action-specific models separately for each common action, by taking one action from couple 2 as train set and the related one from couple 1 as test set.

Unseen action split. The train set is the entire set of common actions $\{\mathcal{A}_c^1, \mathcal{A}_c^2\}$. We regard the extra couple-specific actions $\{\mathcal{A}_u^1, \mathcal{A}_u^2\}$ as unseen actions and use them as our test set. Thus the train and test data contain both couples of actors, but the test actions are not used in training.

To sum up, the common action split is designed for a single model on different actions, the single action split is designed for action-wise models, and the unseen action split focuses on testing unseen actions to measure methods generalization.

5.6.2 EVALUATION METRICS

The most common metric for evaluating 3D joint position in pose estimation and motion prediction tasks is the mean per joint position error MPJPE$(P, G) = \frac{1}{J}\sum_{j=1}^{J} \|P_j - G_j\|_2$, where J is the number of joints, P_j and G_j are the estimated and ground truth position of joint j. Based on MPJPE, we propose two different metrics to evaluate the multi-person motion task.

Joint mean error (JME): We Propose *Joint Mean per joint position Error* to measure poses of different persons in the same coordinate and denote it as JME for simplicity:

$$\text{JME}(P, G) = \text{MPJPE}(P, G), \tag{5.5}$$

where P and G are normalized (see Section 5.5.3) prediction and ground truth. JME provides an overall idea for the performance of collaborative motion prediction by considering the two interacted persons jointly as a whole, measuring both the error of poses and the error of their relative position.

Aligned mean error (AME): We propose *Aligned Mean per joint position Error* to measure pure pose error without the position bias. We first erase the errors on the relative position between the two persons by normalizing the poses independently to obtain $\hat{P}, \hat{G}$. However, the precision of $\hat{P}$ is importantly influenced by the joints that are used to determine the coordinate (hips and back). To mitigate this effect, we compute the best rigid alignment T_A between the estimated pose and the ground-truth using Procrustes analysis [54]:

$$\text{AME}(P, G) = \text{MPJPE}(T_A(\hat{P}, \hat{G}), \hat{G}), \tag{5.6}$$

where $\hat{P} \in [\hat{P}^\ell, \hat{P}^f]$ are independently normalized predictions $\hat{P}^\ell_t = T_{P^\ell_t} P^\ell_t$ and $\hat{P}^f_t = T_{P^f_t} P^f_t$, and T_P is the normalisation transformation computed from the pose P as defined in Section 5.5.3. The same calculation is done for the ground truth $\hat{G}$. This normalization is only used for evaluation purposes.

5.6.3 IMPLEMENTATION DETAILS

Since this is the first time the collaborative motion prediction task is presented in the literature, there are no available methods to compare with. Thus we choose 4 code-released state-of-the-art methods of single-person motion prediction [114, 113, 111, 42], and implement their released codes[5] on ExPI dataset. For a fair comparison, all these models are trained with 50 frames of input, train/test for the leader and the follower separately.

---

[5] All the codes we use are under MIT license.

We train our model for 25 epochs and calculate the average MPJPE loss of 10 predicted frames. As the data is normalized by the leader, the corresponding branch converges faster, so we compensate by exponentially down-weighting the loss of the leader with the number of epochs ϵ, using the loss function: $\mathcal{L} = \mathcal{L}_f + 10^{-\epsilon}\mathcal{L}_{l}$,.

When predicting longer horizons, we use the predicted motion as input to predict future motion. Inspired by [111], we take 64 sub-sequences for each sequence to reduce the variance of the test results. Overall, we have $7k$ and $2.3k$ sub-sequences for training and testing respectively in the common action split and the single action split, and $12k$ / $2.9k$ training/testing samples in the unseen action split.

5.6.4 RESULTS AND DISCUSSION

Common action split. Table 5.3 reports the results of the common action split. We observe that our proposed method outperforms other methods systematically almost for all actions, in all metrics, and for different testing times. In Figure 5.10-left we calculate the percentage of improvement of our method compared with the state-of-the-art methods, and find that we significantly surpass these methods up to $10 \sim 40\%$ on the short term and $5 \sim 30\%$ on the long term. We further compare our per-joint results with Hisrep [111] and MSR [42] in Figure 5.10-right, and observe that our proposed method gets better results on almost on all the joints. More importantly, the keypoints of the limbs (joints of arms and legs) are improved largely. This is reasonable as the interaction between persons comes mostly through the limbs, while joints on the torso have little influence on it. So our cross-interaction attention is able to improve the accuracy on the limbs more than on the torso. We could also notice the large improvement on the feet of the follower which usually flies in the air, indicating that our method works even better for these extremely high dynamic joints.

Single action split and unseen action split. We also reported our proposed method by reporting the results on single action split and unseen action split. For single action split, XIA outperforms the state-of-the-art methods on action-specific models, as shown in Table 5.4. Interestingly, we observe that the performance on the single action split is

Table 5.3: Results on common action split with the two evaluation metrics (in mm). Lower value means better performance. Obviously, our proposal outperforms all the other methods both on JME and AME.

	Action	A1 A-frame				A2 Around the back				A3 Coochie				A4 Frog classic			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [114]	83	141	182	236	127	224	305	433	99	177	239	350	74	135	182	250
	LTD [113]	70	125	157	189	131	242	321	426	102	194	260	357	62	117	155	197
	Hisrep [111]	52	103	139	188	96	186	256	349	57	118	167	240	45	93	131	180
	MSR [42]	56	100	**132**	**175**	102	187	256	365	65	120	166	244	50	95	127	172
	Ours	**49**	**98**	140	192	**84**	**166**	**234**	**346**	**51**	**105**	**154**	**234**	**41**	**84**	**120**	**161**
AME	Res-RNN [114]	59	102	132	167	62	112	152	229	57	102	139	215	48	85	113	157
	LTD [113]	51	92	116	132	51	91	116	**148**	43	80	103	130	38	70	89	111
	Hisrep [111]	34	69	97	130	44	84	**115**	150	32	65	91	121	27	56	82	112
	MSR [42]	41	75	99	**126**	54	96	129	180	41	74	98	135	34	61	82	106
	Ours	**32**	**68**	**99**	128	**41**	**82**	116	163	**29**	**58**	**84**	**116**	**24**	**50**	**73**	**96**

	Action	A5 Noser				A6 Toss Out				A7 Cartwheel				AVG			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [114]	87	152	201	271	93	166	225	321	104	189	269	414	95	169	229	325
	LTD [113]	72	131	173	231	81	151	200	280	112	223	315	442	90	169	226	303
	Hisrep [111]	51	105	149	214	61	125	176	252	71	150	222	333	62	126	177	251
	MSR [42]	54	100	138	202	70	132	182	258	82	154	218	321	69	127	174	248
	Ours	**43**	**90**	**132**	**197**	**55**	**113**	**163**	**242**	**62**	**130**	**192**	**291**	**55**	**112**	**162**	**238**
AME	Res-RNN [114]	51	90	120	167	53	94	126	183	74	131	178	265	58	102	137	197
	LTD [113]	39	70	90	116	42	75	94	123	52	101	139	198	45	83	107	137
	Hisrep [111]	28	58	85	121	34	66	88	115	42	83	120	171	34	69	97	131
	MSR [42]	33	59	79	**109**	42	71	93	124	57	103	146	210	43	77	104	141
	Ours	**24**	**51**	**75**	**109**	**31**	**62**	**86**	**114**	**41**	**81**	**115**	**160**	**32**	**65**	**93**	**127**

Table 5.4: Results on single action split with the two evaluation metrics (in mm). Lower value means better performance. Seven action-wise models are trained independently. Our method performs the best in 5 actions, and close to the best for the other 2 actions.

	Action	A1 A-frame				A2 Around the back				A3 Coochie			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [114]	75	131	171	226	122	215	287	403	97	174	235	329
	LTD [113]	70	126	155	183	131	243	312	415	102	194	252	338
	Hisrep [111]	66	118	153	190	128	231	308	417	74	143	205	295
	MSR[42]	**64**	**108**	**136**	**171**	119	210	282	385	79	144	189	265
	Ours	**64**	120	160	199	**109**	**200**	**275**	**381**	**59**	**117**	**174**	**277**
AME	Res-RNN [114]	56	99	129	163	61	110	150	229	53	96	131	188
	LTD [113]	51	93	114	127	**51**	**91**	**116**	**162**	43	80	100	**126**
	Hisrep [111]	45	83	106	118	57	102	135	178	39	72	100	132
	MSR[42]	46	79	**98**	**118**	60	107	141	192	48	86	111	150
	Ours	**43**	**84**	115	131	53	99	136	185	**35**	**68**	**98**	140

	Action	A4 Frog classic				A5 Noser				A6 Toss Out				A7 Cartwheel			
	Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
JME	Res-RNN [114]	73	131	177	246	76	136	184	255	100	184	252	357	88	162	219	293
	LTD [113]	62	117	153	203	71	131	171	231	81	151	199	299	112	223	306	411
	Hisrep [111]	64	120	159	191	63	121	166	227	90	168	232	312	88	166	232	332
	MSR[42]	**59**	**103**	**134**	**173**	65	118	162	225	86	151	201	283	96	178	255	362
	Ours	60	116	162	209	**53**	**106**	**152**	**221**	**65**	**122**	**166**	**223**	**74**	**144**	**203**	**301**
AME	Res-RNN [114]	46	81	106	142	44	79	106	147	53	100	162	176	70	133	163	198
	LTD [113]	38	70	**88**	118	39	70	90	125	42	75	**93**	123	52	101	137	188
	Hisrep [111]	41	77	103	119	35	70	97	125	46	82	107	137	48	90	121	169
	MSR[42]	39	**68**	**88**	**111**	39	69	91	**121**	55	93	117	156	66	118	163	222
	Ours	**37**	74	106	128	**29**	**59**	**86**	125	**39**	**72**	94	**119**	**43**	**82**	**112**	**152**

worse than the corresponding results on the common action split, meaning that training on different actions helps regularise the network for this very challenging collaborative extreme motion prediction task. Regarding the unseen action split shown in Table 5.5, we can see that XIA still outperforms the state-of-the-art methods on most of the actions, demonstrating the robustness of our method.

Qualitative results. Figure 5.1 shows some examples of our visualization results compared to Hisrep *et al.* [111], MSR [42] and the ground truth, on the common action split. More qualitative examples could be found in Figure 5.11, Figure 5.12 and Figure 5.13 where we compare our model with models that independently predict the motion of each person, i.e. Res-RNN [114], LTD [113], Hisrep [111] and MSR [42]. We can see that the poses estimated by our method are much closer to the ground truth than the other methods, and it works well even on some extreme actions where other methods totally fail (Figure 5.1-right).

Ablation study. Taking Hisrep [111] as an example, we first tried 3 different ways of training the single-person motion prediction models on our multi-person dataset: (i) 'mix': train a single model using data of the two poses $\{P^l, P^f\}$; (ii) 'cat': concatenate the two poses as a single input vector $[P^l, P^f]$; (iii) 'sep': train two person-specific models for P^l and P^f. Since 'sep' gives the best performance, all the state-of-the-art methods reported above in this chapter is using this setting. As for our collaborative motion prediction model, we report the performances of several different design choices of our model. We found that updating the key and values of the temporal attention using our XIA module provides the best results. We demonstrate interest in the design of our method as the proposed one is the best in performance and our method significantly improves all the single-person motion prediction results.

5.7 CONCLUSION

Current motion prediction methods are restricted to a single person. In this chapter, we move beyond existing approaches for 3D human motion prediction by considering

Table 5.5: Action-wise results on unseen action split with the two evaluation metrics (in mm). Lower value means better performance. Our method still performs the best on most of the unseen actions and on the average result.

	Action	A8			A9			A10			A11			A12		
	Time (sec)	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8
JME	Res-RNN [114]	239	312	371	193	256	303	189	257	310	305	425	520	215	289	348
	LTD [113]	239	324	394	175	226	259	148	191	220	176	240	286	143	178	192
	Hisrep [111]	195	**283**	**358**	121	169	206	92	129	**160**	129	193	245	**80**	**104**	121
	MSR [42]	230	289	335	188	245	290	148	198	248	234	319	384	176	232	278
	Ours	**191**	287	377	**118**	**165**	**203**	**91**	**129**	162	**122**	**183**	**232**	81	107	**128**
AME	Res-RNN. [114]	124	165	195	125	157	181	131	166	189	148	198	240	149	169	192
	LTD [113]	**95**	**123**	**146**	85	106	116	74	91	101	86	115	137	98	125	134
	Hisrep [111]	101	144	176	61	82	94	**49**	**67**	**80**	73	**105**	**129**	**53**	**73**	**86**
	MSR [42]	103	134	155	101	135	160	74	98	121	103	143	173	87	111	132
	Ours	**95**	137	171	**58**	**80**	**93**	51	70	84	**70**	**105**	134	**53**	**73**	88

	Action	A13			A14			A15			A16			AVG		
	Time (sec)	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8
JME	Res-RNN [114]	165	214	252	214	293	357	149	187	210	167	226	277	204	273	327
	LTD [113]	146	193	226	252	333	387	174	228	264	139	184	217	177	233	272
	Hisrep [111]	112	154	187	157	219	257	134	190	233	**96**	**146**	**187**	124	176	**218**
	MSR [42]	162	218	266	177	239	295	143	179	213	157	222	281	179	238	288
	Ours	**106**	**150**	**185**	**156**	**216**	**256**	**126**	**175**	**213**	**96**	152	205	**121**	**174**	**218**
AME	Res-RNN. [114]	102	128	147	181	237	279	100	129	144	93	124	147	128	164	190
	LTD [113]	85	110	124	106	136	155	91	119	135	72	96	116	88	113	129
	Hisrep [111]	64	89	**104**	86	120	**142**	73	104	128	54	82	**104**	68	96	**116**
	MSR [42]	84	106	122	88	118	**142**	90	113	136	90	122	148	91	120	143
	Ours	**63**	**88**	**104**	**82**	**116**	**142**	**69**	**97**	**120**	**52**	**79**	**104**	**66**	**94**	**116**

Table 5.6: Ablations. *'mix /cat /sep'* use the single person motion prediction model (Hisrep [111]) for multi-person by: mixing two poses together / concatenate two poses as a single vector / train two person-specific models. *'w.o. XIA'* indicates training leader and follower in parallel using our defined loss without XIA module; *'XIA kqv / kq / kv / v'* use XIA module to update key, value and query of the temporal attention, or just some of them.

	JME					**AME**				
Time (sec)	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
mix	69	132	185	233	271	41	77	104	126	142
cat	61	123	176	223	262	37	71	99	121	138
sep	62	126	177	218	251	34	69	97	116	131
w.o. XIA	58	120	174	217	249	33	68	98	118	131
XIA kq	58	118	169	211	245	33	67	95	114	128
XIA kqv	57	117	170	215	251	32	65	95	116	131
XIA v	56	116	168	210	244	32	66	94	113	**127**
XIA kv	**55**	**112**	**162**	**204**	**238**	**32**	**65**	**93**	**112**	**127**

a scenario with two persons performing highly interactive activities. We collected a new dataset called ExPI of professional actors performing dancing actions. ExPI is annotated with sequences of 3D body poses and shapes, opening the door to not only being applied for interactive motion prediction but also for single-frame pose estimation or multi-view 3D reconstruction. In order to learn the interacted motion dynamics, we introduce a baseline method trained with ExPI that exploits historical information of both people in an attention-like fashion. The results of our method show consistent improvement compared to methods that independently predict the motion of each person.

While collecting clean and reusable 3D pose data requires specific equipment and recording extreme poses requires actors with specific skills, thus ExPI is rare and difficult to reproduce/extend. This is clearly a limitation in the era of data-hungry deep learning architectures. Besides, we observe that most current methods designed for human motion prediction, either previous methods for single-person or our proposed method for multi-person, are all based on complex architectures. The complexity not only means complex architecture designs but also big model sizes. Thus we begin to think if we could have a simpler design to deal with human motion data with fewer model parameters. This is the motivation for our next chapter.

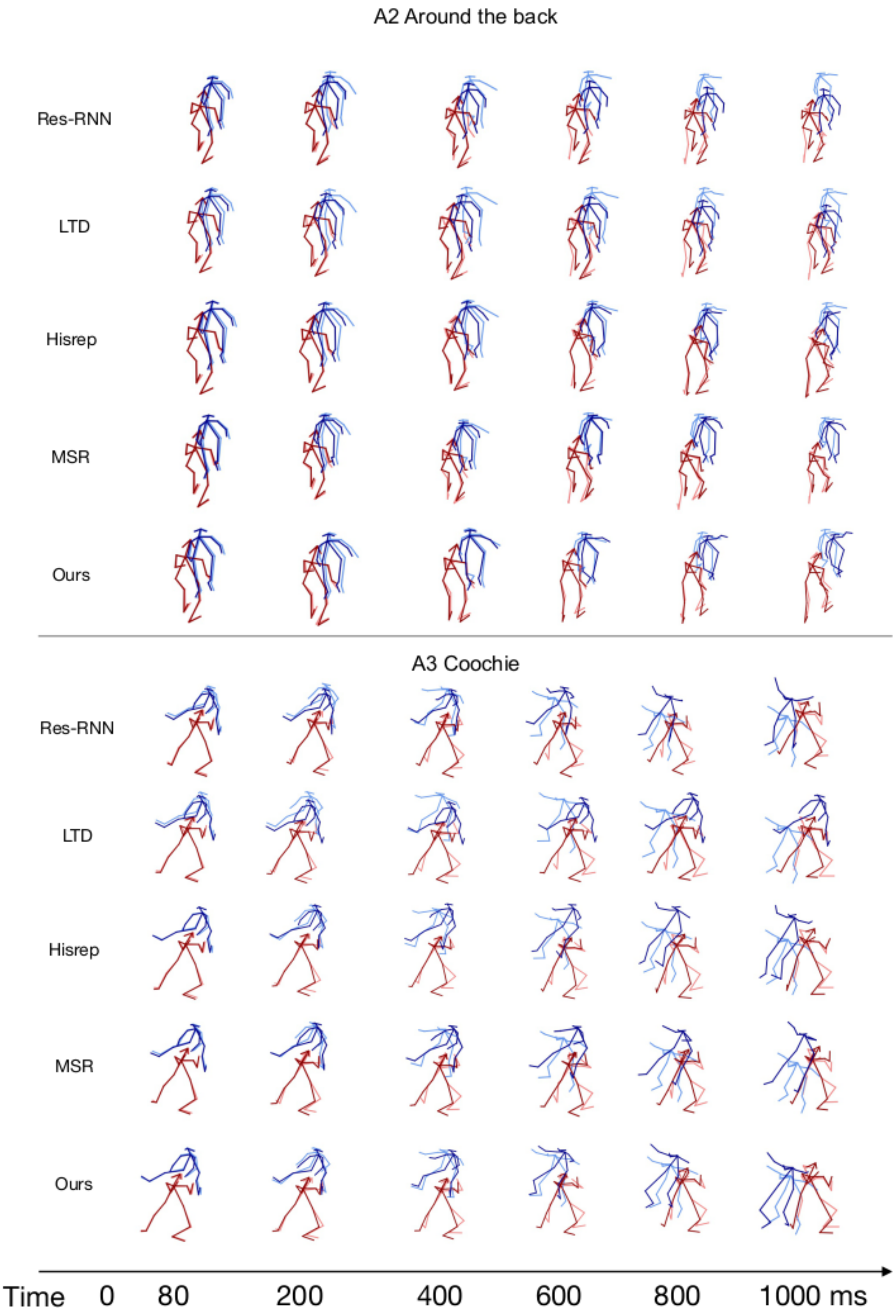

A2 Around the back
Res-RNN
LTD
Hisrep
MSR
Ours
A3 Coochie
Res-RNN
LTD
Hisrep
MSR
Ours
Time
0
80
200
400
600
800
1000 ms

Figure 5.11: Visualization results of our proposed method compared with previous state-of-the-art methods. Dark red/blue shows predicted results, and light red/blue represents groundtruth.

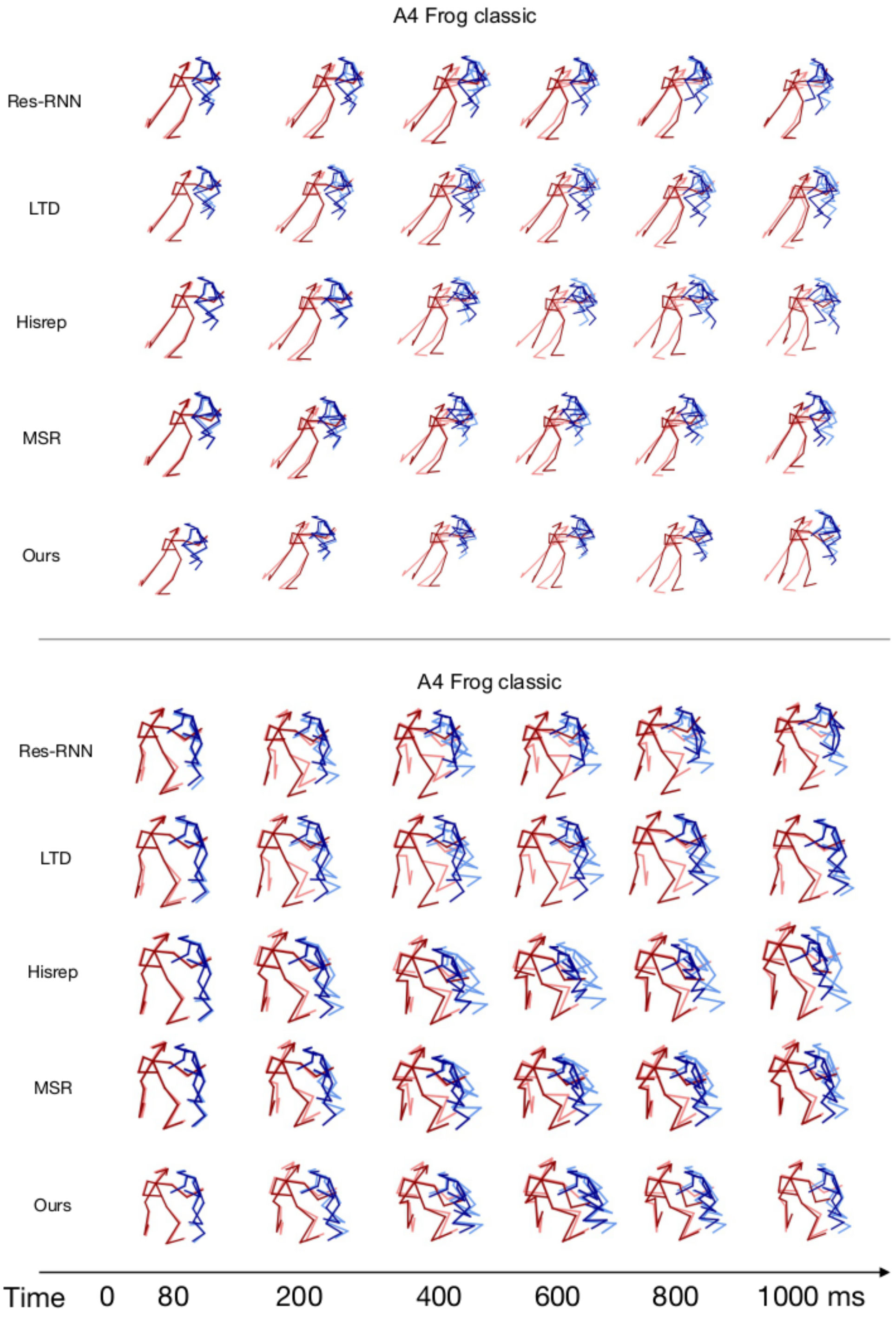

A4 Frog classic
Res-RNN
LTD
Hisrep
MSR
Ours
A4 Frog classic
Res-RNN
LTD
Hisrep
MSR
Ours
Time 0 80 200 400 600 800 1000 ms

Figure 5.12: Visualization results of our proposed method compared with previous state-of-the-art methods (continue). Dark red/blue shows predicted results, and light red/blue represents groundtruth.

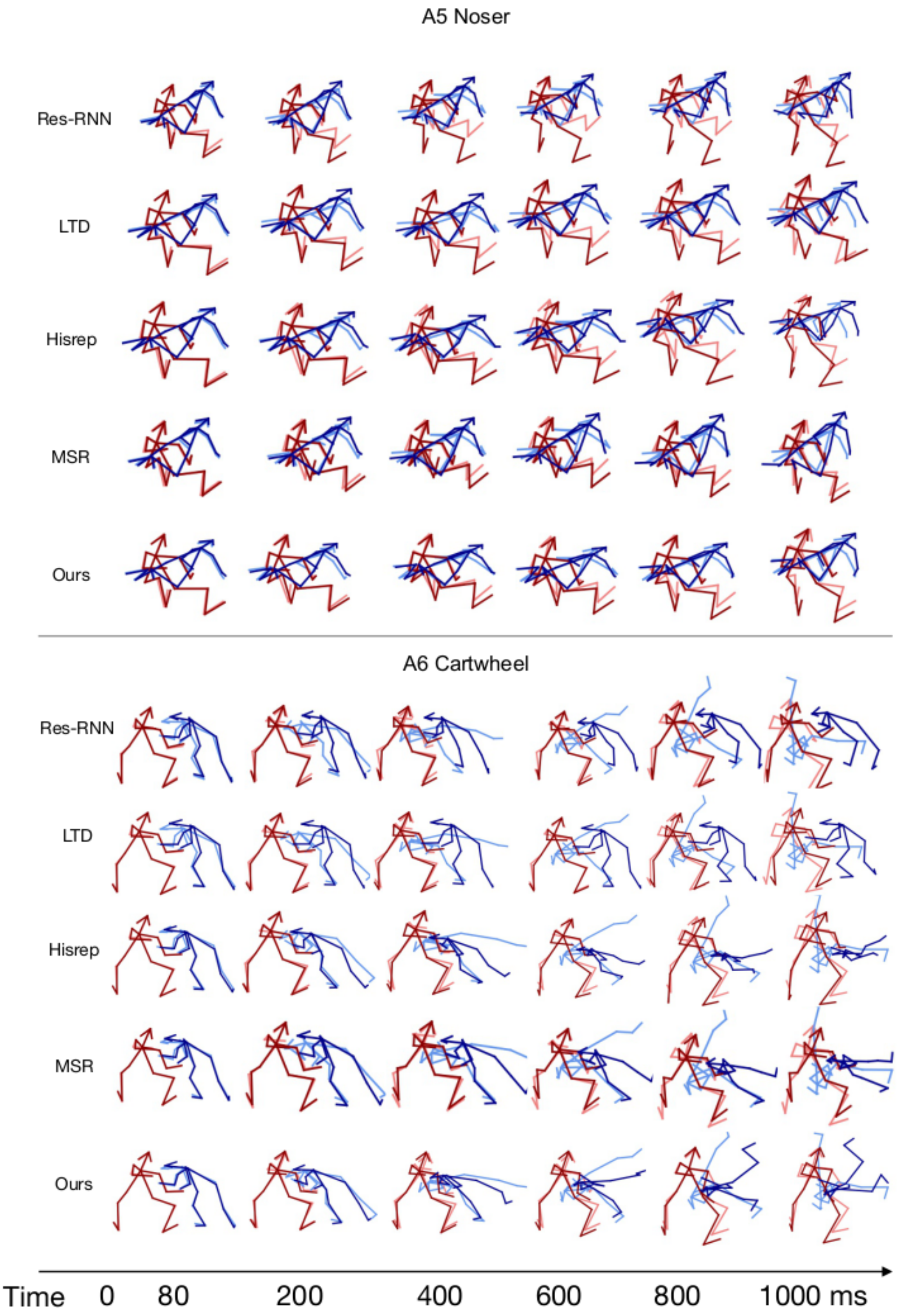

A5 Noser
Res-RNN
LTD
Hisrep
MSR
Ours
A6 Cartwheel
Res-RNN
LTD
Hisrep
MSR
Ours
Time
0
80
200
400
600
800
1000 ms

Figure 5.13: Visualization results of our proposed method compared with previous state-of-the-art methods (continue). Dark red/blue shows predicted results, and light red/blue represents groundtruth.

CHAPTER 6

A SIMPLE BASELINE FOR HUMAN MOTION PREDICTION

This chapter tackles the problem of human motion prediction, consisting in forecasting future body poses from historically observed sequences. State-of-the-art approaches provide good results, however, they rely on deep learning architectures of arbitrary complexity, such as Recurrent Neural Networks (RNN), Transformers, or Graph Convolutional Networks (GCN), typically requiring multiple training stages and more than 2 million parameters. In this chapter, we first show the effectiveness of a single fully connected layer on the motion prediction task through two naive experiments, and thus propose siMLPe, a simple yet effective network that is composed of merely three components: the fully connected layers, the layer normalization, and the matrix transpose operations. The network has in total 0.14 million parameters which are about 30 times smaller than the state-of-the-art methods, and it has achieved state-of-the-art results on multiple benchmarks.

The results for single-person human motion prediction presented in this chapter were initially presented in: “Back to MLP: A simple baseline for human motion prediction”, Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer, In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023.

We also present in this chapter some extended experiments on multi-person settings. The code of this work is released at `https://github.com/dulucas/siMLPe`.

6.1 INTRODUCTION

As presented in the previous chapter, the goal of human motion prediction is to forecast the follow-up of a sequence of 3D body poses. Accurate prediction of future human motion is vital for several applications, such as accident prevention in autonomous driving [128], people tracking [53], and human-robot interaction [82].

Due to the spatio-temporal nature of human motion, the common trend in the literature is to design models that are capable of fusing spatial and temporal information. Traditional approaches mainly relied on hidden Markov models [23] or Gaussian process latent variable models [166]. However, while these approaches performed well on sim-

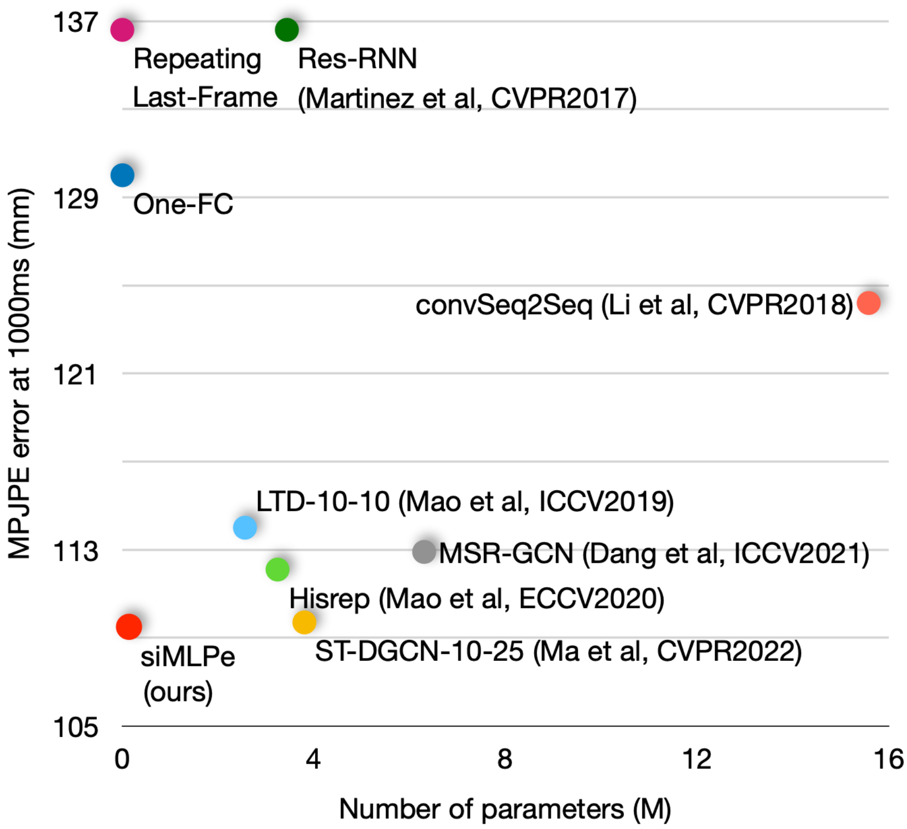


Figure 6.1: **Comparison of parameter size and performance on the Human3.6M dataset [75].** We report the MPJPE metric in *mm* at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. Our method (SIMLPE, in red) achieves the lowest error with significantly fewer parameters. We also show the performance of two simple methods: 'Repeating Last-Frame' systematically repeats the last input frame as output prediction, and 'One-FC' uses only one single fully connected layer to predict the future motion.

ple and periodic motion patterns, they dramatically fail under complex motions [113]. In recent years, with the success of deep learning, various methods have been developed based on different types of neural networks that are able to handle sequential data. For example, some works use Recurrent Neural Networks (RNN) [114] to model the human motion [49, 76, 114, 103, 33], and some more recent works [113, 111, 60, 107, 41, 96, 94] propose networks based on Graph Convolutional Networks (GCN) [113], or trying with Transformers([5]) based method [111, 5, 26] to fuse the spatial and temporal information of the motion sequence across human joints and time. However, the architectures of these recent methods are usually not simple and some of them require additional priors, which makes their network difficult to analyze and modify. Thus, a question naturally arises: *"Can we tackle the human motion prediction with a simple network?"*

To answer this question, we first tried a naive solution by just repeating the last input pose and using it as the output prediction. As shown in Figure 6.1, this naive solution could already achieve reasonable results, which means the last input pose is "close" to the future poses. Inspired by this, we further train only one fully connected layer to predict the residual between the future poses and the last input pose and achieve better performance, which shows the potential of a simple network for human motion prediction built on basic layers like the fully connected layer.

Based on the above observations, we go back to the multi-layer perceptrons (MLPs) and build a simple yet effective network named SIMLPE with only three components: fully connected layers, layer normalization [13], and transpose operations. The network architecture is shown in Figure 6.2. Noticeably, we found that even commonly used activation layers such as ReLU [125] are not needed, which makes our network an entirely linear model except for layer normalization. Despite its simplicity, SIMLPE achieves strong performance when appropriately combined with three simple practices: applying the Discrete Cosine Transform (DCT), predicting residual displacement of joints, and optimizing velocity as an auxiliary loss.

SIMLPE yields state-of-the-art performance on several standard benchmarks, including Human3.6M [75], AMASS [109] and 3DPW [163]. In the meantime, SIMLPE is lightweight and requires $20\times$ to $60\times$ fewer parameters than previous state-of-the-art approaches. A comparison between SIMLPE and previous methods can be found in Figure 6.1, which shows the Mean Per Joint Position Error (MPJPE) at $1,000$ms on Human3.6M of different networks versus the network complexity. SIMLPE achieves the best performance with high efficiency.

In summary, our contributions are as follows:

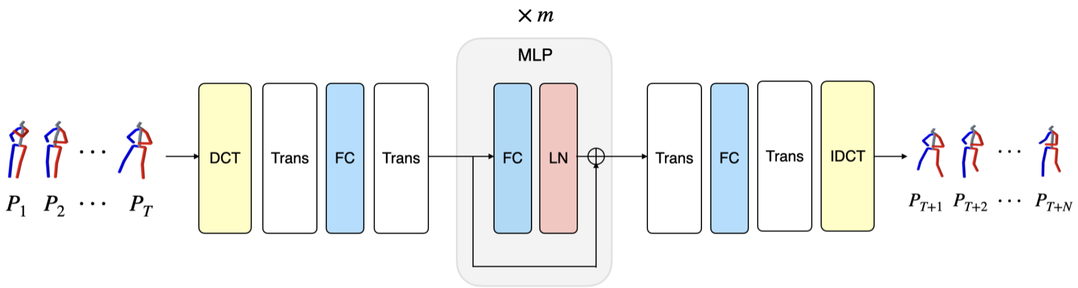


Figure 6.2: **Overview of our approach siMLPe for human motion prediction.** *FC* denotes a fully connected layer, *LN* denotes layer normalization [13], and *Trans* represents the transpose operation. *DCT* and *IDCT* represent the discrete cosine transformation and inverse discrete cosine transformations respectively. The MLP blocks (in gray), composing FC and LN, are repeated m times.

- We show that human motion prediction can be modeled in a simple way without explicitly fusing spatial and temporal information. As an extreme example, a single fully connected layer can already achieve reasonable performance.
- We propose siMLPe, a simple yet effective network for human motion prediction with only three components: fully connected layers, layer normalization, and transpose operation, achieving state-of-the-art performance with far fewer parameters than existing methods on multiple benchmarks such as Human3.6M, AMASS and 3DPW datasets.

6.2 RELATED WORK

In Section 2.2 we discussed the previous literature on single-person human motion prediction, based on RNN, GCN, or Attention mechanisms. In summary, with the development of human motion prediction in recent years, the RNN/GCN/transformer-based architectures are well explored and the results have been significantly improved. Though these methods provide good results, their architectures are becoming more and more complicated and difficult to train. Especially after LTD [113] was proposed, the most recent works [86, 107, 41, 96, 94] focus on designing complex networks based on GCN, to model the spatial and temporal information at different level. While based on the observation that the human motion data is simple, we doubt the necessity of designing a complicated network for this task. In this chapter, we stick to simple architectures and propose an MLP-based network. Recently, a concurrent and independent work [22] based on [154] also adopts an MLP-based network architecture for motion prediction, while our network is much simpler as we do not use the squeeze-and-excitation block[72] nor the activation layers. We hope that our simple method will serve as a baseline and let the community rethink the problem of human motion prediction.

6.3 OUR APPROACH: SIMLPE

In this section, we formulate the problem and present the formulation of the DCT transformation in Section 6.3.1, details of the network architecture in Section 6.3.2, and the losses we use for training in Section 6.3.3.

Given a sequence of 3D human poses in the past, our goal is to predict the future sequence of poses. We denote the observed 3D human poses as $\mathbf{P}_{1:T} = [P_1^\top, .., P_T^\top]^\top \in \mathbb{R}^{T\times C}$, consisting of T consecutive human poses, where the pose at the t-th frame P_t is represented by a C-dimensional vector, i.e. $P_t \in \mathbb{R}^C$. In this work, similar to previous works [114, 113, 111, 107], P_t is the 3D coordinates of joints at t-th frame and $C = 3 \times \mathbf{J}$, where $\mathbf{J}$ is the number of joints. Our task is to predict the future N motion frames $\mathbf{G}_{T+1:T+N} = [P_{T+1}^\top, .., P_{T+N}^\top]^\top \in \mathbb{R}^{N\times C}$.

6.3.1 DISCRETE COSINE TRANSFORM (DCT)

We adopt the DCT transformation to encode temporal information, which is proven to be beneficial for human motion prediction [113, 111, 107]. More precisely, given an input

Table 6.1: **Results on Human3.6M** for different prediction time steps (ms). We report the MPJPE error in *mm* and number of parameters (M) for each method. Lower is better. 256 samples are tested for each action. † indicates that the results are taken from the paper [111], ⋆ indicated that the results are taken from the paper [107]. Note that ST-DGCN [107] use two different models to evaluate their short-/long- term performance, here we report their results of a single model which performs better on long-term for fair comparison. We also show results of two simple baselines: 'Repeating Last-Frame' repeats the last input frame 25 times as output, 'One FC' uses only one single fully connected layer for the prediction.

	MPJPE (mm) ↓								# Param.(M) ↓
Time (ms)	80	160	320	400	560	720	880	1000	
Repeating Last-Frame	23.8	44.4	76.1	88.2	107.4	121.6	131.6	136.6	0
One FC	14.0	33.2	68.0	81.5	101.7	115.1	124.8	130.0	0.003
Res-RNN † [114]	25.0	46.2	77.0	88.3	106.3	119.4	130.0	136.6	3.44
convSeq2Seq † [89]	16.6	33.3	61.4	72.7	90.7	104.7	116.7	124.2	15.58
LTD-50-25 † [113]	12.2	25.4	50.7	61.5	79.6	93.6	105.2	112.4	2.56
LTD-10-10 † [113]	11.2	23.4	47.9	58.9	78.3	93.3	106.0	114.0	2.55
Hisrep † [111]	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1	3.24
MSR-GCN ⋆ [41]	11.3	24.3	50.8	61.9	80.0	-	-	112.9	6.30
ST-DGCN-10-25 ⋆ [107]	10.6	23.1	47.1	57.9	76.3	90.7	102.4	109.7	3.80
SIMLPE (Ours)	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**	**0.14**

motion sequence of T frames, the DCT matrix $\mathbf{D} \in \mathbb{R}^{T\times T}$ can be calculated as:

$$\mathbf{D}_{i,j} = \sqrt{\frac{2}{T}} \frac{1}{\sqrt{1+\delta_{i,0}}} \cos\left(\frac{\pi}{2T}(2j+1)i\right) , \tag{6.1}$$

where $\delta_{i,j}$ denotes the *Kronecker* delta:

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \, . \end{cases} \tag{6.2}$$

The transformed input is $\mathcal{D}(\mathbf{P}_{1:T}) = \mathbf{D}\mathbf{P}_{1:T}$. We apply the Inverse Discrete Cosine Transform (IDCT) to transform the output of the network back to the original pose representation, denoted as $\mathcal{D}^{-1}$ and the inverse of $\mathbf{D}$.

Table 6.2: **Action-wise results on Human3.6M** for different prediction time steps (ms). Lower is better. 256 samples are tested for each action. † indicates that the results are taken from the paper [111], ⋆ indicates that the results are taken from the paper [107].

Action	walking				eating				smoking				discussion			
Time (ms)	80	400	560	1000	80	400	560	1000	80	400	560	1000	80	400	560	1000
Res-RNN † [114]	23.2	66.1	71.6	79.1	16.8	61.7	74.9	98.0	18.9	65.4	78.1	102.1	25.7	91.3	109.5	131.8
convSeq2Seq † [89]	17.7	63.6	72.2	82.3	11.0	48.4	61.3	87.1	11.6	48.9	60.0	81.7	17.1	77.6	98.1	129.3
LTD-50-25 † [113]	12.3	44.4	50.7	60.3	7.8	38.6	51.5	75.8	8.2	39.5	50.5	72.1	11.9	68.1	88.9	118.5
LTD-10-10 † [113]	11.1	42.9	53.1	70.7	7.0	37.3	51.1	78.6	7.5	37.5	49.4	71.8	10.8	65.8	88.1	121.6
Hisrep † [111]	10.0	39.8	47.4	58.1	6.4	36.2	50.0	75.7	7.0	36.4	47.6	69.5	10.2	65.4	86.6	119.8
MSR-GCN ⋆ [41]	10.8	42.4	53.3	63.7	6.9	36.0	50.8	75.4	7.5	37.5	50.5	72.1	10.4	65.0	87.0	116.8
ST-DGCN-10-25 ⋆ [107]	11.2	42.8	49.6	58.9	6.5	36.8	50.0	74.9	7.3	37.5	48.8	69.9	10.2	64.4	86.1	116.9
siMLPe (Ours)	**9.9**	**39.6**	**46.8**	**55.7**	**5.9**	**36.1**	**49.6**	**74.5**	**6.5**	**36.3**	**47.2**	**69.3**	**9.4**	**64.3**	**85.7**	**116.3**
Action	directions				greeting				phoning				posing			
Time (ms)	80	400	560	1000	80	400	560	1000	80	400	560	1000	80	400	560	1000
Res-RNN † [114]	21.6	84.1	101.1	129.1	31.2	108.8	126.1	153.9	21.1	76.4	94.0	126.4	29.3	114.3	140.3	183.2
convSeq2Seq † [89]	13.5	69.7	86.6	115.8	22.0	96.0	116.9	147.3	13.5	59.9	77.1	114.0	16.9	92.9	122.5	187.4
LTD-50-25 † [113]	8.8	58.0	74.2	105.5	16.2	82.6	104.8	136.8	9.8	50.8	68.8	105.1	12.2	79.9	110.2	174.8
LTD-10-10 † [113]	8.0	54.9	76.1	108.8	14.8	79.7	104.3	140.2	9.3	49.7	68.7	105.1	10.9	75.9	109.9	171.7
Hisrep † [111]	7.4	56.5	73.9	106.5	13.7	78.1	101.9	138.8	8.6	49.2	67.4	105.0	10.2	75.8	107.6	178.2
MSR-GCN ⋆ [41]	7.7	56.2	75.8	105.9	15.1	85.4	106.3	**136.3**	9.1	49.8	67.9	104.7	10.3	75.9	112.5	176.5
ST-DGCN-10-25 ⋆ [107]	7.5	56.0	73.3	**105.9**	14.0	77.3	100.2	136.4	8.7	48.8	66.5	**102.7**	10.2	73.3	102.8	**167.0**
siMLPe (Ours)	**6.5**	**55.8**	**73.1**	106.7	**12.4**	**77.3**	**99.8**	137.5	**8.1**	**48.6**	**66.3**	103.3	**8.8**	**73.8**	**103.4**	168.7
Action	purchases				sitting				sittingdown				takingphoto			
Time (ms)	80	400	560	1000	80	400	560	1000	80	400	560	1000	80	400	560	1000
Res-RNN † [114]	28.7	100.7	122.1	154.0	23.8	91.2	113.7	152.6	31.7	112.0	138.8	187.4	21.9	87.6	110.6	153.9
convSeq2Seq † [89]	20.3	89.9	111.3	151.5	13.5	63.1	82.4	120.7	20.7	82.7	106.5	150.3	12.7	63.6	84.4	128.1
LTD-50-25 † [113]	15.2	78.1	99.2	134.9	10.4	58.3	79.2	118.7	17.1	76.4	100.2	143.8	9.6	54.3	75.3	118.8
LTD-10-10 † [113]	13.9	75.9	99.4	135.9	9.8	55.9	78.5	118.8	15.6	71.7	96.2	142.2	8.9	51.7	72.5	116.3
Hisrep † [111]	13.0	73.9	95.6	134.2	9.3	56.0	76.4	115.9	14.9	72.0	97.0	143.6	8.3	51.5	72.1	115.9
MSR-GCN ⋆ [41]	13.3	77.8	99.2	134.5	9.8	55.5	77.6	115.9	15.4	73.8	102.4	149.4	8.9	54.4	77.7	121.9
ST-DGCN-10-25 ⋆ [107]	13.2	74.0	95.7	**132.1**	9.1	54.6	75.1	114.8	14.7	**70.0**	**94.4**	**139.0**	8.2	**50.2**	**70.5**	112.9
siMLPe (Ours)	**11.7**	**72.4**	**93.8**	132.5	**8.6**	**55.2**	**75.4**	**114.1**	**13.6**	70.8	95.7	142.4	**7.8**	50.8	71.0	**112.8**
Action	waiting				walkingdog				walkingtogether				average			
Time (ms)	80	400	560	1000	80	400	560	1000	80	400	560	1000	80	400	560	1000
Res-RNN † [114]	23.8	87.7	105.4	135.4	36.4	110.6	128.7	164.5	20.4	67.3	80.2	98.2	25.0	88.3	106.3	136.6
convSeq2Seq † [89]	14.6	68.7	87.3	117.7	27.7	103.3	122.4	162.4	15.3	61.2	72.0	87.4	16.6	72.7	90.7	124.2
LTD-50-25 † [113]	10.4	59.2	77.2	108.3	22.8	88.7	107.8	156.4	10.3	46.3	56.0	65.7	12.2	61.5	79.6	112.4
LTD-10-10 † [113]	9.2	54.4	73.4	107.5	20.9	86.6	109.7	150.1	9.6	44.0	55.7	69.8	11.2	58.9	78.3	114.0
Hisrep † [111]	8.7	54.9	74.5	108.2	20.1	86.3	108.2	146.9	8.9	41.9	52.7	64.9	10.4	58.3	77.3	112.1
MSR-GCN ⋆ [41]	10.4	62.4	74.8	105.5	24.9	112.9	107.7	145.7	9.2	43.2	56.2	69.5	11.3	61.9	80.0	112.9
ST-DGCN-10-25 ⋆ [107]	8.7	53.6	71.6	**103.7**	20.4	84.6	105.7	145.9	8.9	43.8	54.4	64.6	10.6	57.9	76.3	109.7
siMLPe (Ours)	**7.8**	**53.2**	**71.6**	104.6	**18.2**	**83.6**	**105.6**	**141.2**	**8.4**	**41.2**	**50.8**	**61.5**	**9.6**	**57.3**	**75.7**	**109.4**

6.3.2 NETWORK ARCHITECTURE

Figure 6.2 shows the architecture of our network. Our network only contains three components: fully connected layers, transpose operation, and layer normalization [13]. For all the fully connected layers, their input dimension is equal to their output dimension.

Formally, given an input sequence of 3D human poses $\mathbf{P}_{1:T} = [P_1^\top, .., P_T^\top]^\top \in \mathbb{R}^{T \times C}$, our network predicts a sequence of future poses $\mathbf{P}_{T+1:T+N} = [{P'}_{T+1}^\top, .., {P'}_{T+N}^\top]^\top \in \mathbb{R}^{N \times C}$:

$$\mathbf{P}_{T+1:T+N} = \mathcal{D}^{-1}(\mathcal{F}(\mathcal{D}(\mathbf{P}_{1:T}))) \, , \tag{6.3}$$

where $\mathcal{F}$ denotes our network.

After the DCT transformation, we apply one fully connected layer to operate only on the spatial dimension of the transformed motion sequence $\mathcal{D}(\mathbf{P}_{1:T}) \in \mathbb{R}^{T \times C}$:

$$\mathbf{z}^0 = \mathcal{D}(\mathbf{P}_{1:T})\mathbf{W}_0 + \mathbf{b}_0 \, , \tag{6.4}$$

where $\mathbf{z}^0 \in \mathbb{R}^{T \times C}$ is the output of the fully connected layer, $\mathbf{W}_0 \in \mathbb{R}^{C \times C}$ and $\mathbf{b}_0 \in \mathbb{R}^C$ represent the learnable parameters of the fully connected layer. In practice, this is equivalent to applying a transpose operation with a fully connected layer, and then transposing back the output feature, as shown in Figure 6.2.

Then, a series of m blocks are introduced to only operate on the temporal dimension, i.e., only to merge information across frames. Each block consists of a fully connected layer followed by layer normalization, formally:

$$\mathbf{z}^i = \mathbf{z}^{i-1} + \text{LN}(\mathbf{W}_i \mathbf{z}^{i-1} + \mathbf{b}_i) \, , \tag{6.5}$$

where $\mathbf{z}^i \in \mathbb{R}^{T \times C}, i \in [1, .., m]$ denotes the output of the i-th MLP block, LN denotes the layer normalization operation, and $\mathbf{W}_i \in \mathbb{R}^{T \times T}$ and $\mathbf{b}_i \in \mathbb{R}^T$ are the learnable parameters of the fully connected layer in the i-th MLP block.

Finally, similar to the first fully connected layer, we add another fully connected layer after the MLP blocks to operate only on the spatial dimension of the feature, and then apply IDCT transformation to obtain the prediction:

$$\mathbf{P}_{T+1:T+N} = \mathcal{D}^{-1}(\mathbf{z}'\mathbf{W}_{m+1} + \mathbf{b}_{m+1}) \, , \tag{6.6}$$

where $\mathbf{W}_{m+1}$ and $\mathbf{b}_{m+1}$ are the learnable parameters of the last fully connected layer.

Table 6.3: **Results on AMASS and 3DPW** for different prediction time steps (ms). We report the MPJPE error in *mm*. Lower is better. The model is trained on the AMASS dataset. The results of the previous methods are taken from [111].

Dataset	AMASS-BMLrub								3DPW							
Time (ms)	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
convSeq2Seq [89]	20.6	36.9	59.7	67.6	79.0	87.0	91.5	93.5	18.8	32.9	52.0	58.8	69.4	77.0	83.6	87.8
LTD-10-10 [113]	**10.3**	**19.3**	36.6	44.6	61.5	75.9	86.2	91.2	**12.0**	**22.0**	38.9	46.2	59.1	69.1	76.5	81.1
LTD-10-25 [113]	11.0	20.7	37.8	45.3	57.2	65.7	71.3	75.2	12.6	23.2	39.7	46.6	57.9	65.8	71.5	75.5
Hisrep [111]	11.3	20.7	35.7	42.0	51.7	58.6	63.4	67.2	12.6	23.1	39.0	45.4	56.0	63.6	69.7	73.7
siMLPe (Ours)	10.8	19.6	**34.3**	**40.5**	**50.5**	**57.3**	**62.4**	**65.7**	12.1	22.1	**38.1**	**44.5**	**54.9**	**62.4**	**68.2**	**72.2**

Table 6.4: **Average results for different prediction time periods on Human3.6M and AMASS.** These results are obtained following the evaluation method of STS-GCN [146] and STG-GCN [179], instead of the standard evaluation protocol adopted in [113, 111, 107].

Dataset	Human3.6M								AMASS-BMLrub							
Time (ms)	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
STS-GCN [146]	10.1	17.1	33.1	38.3	50.8	60.1	68.9	75.6	10.0	12.5	21.8	24.5	31.9	38.1	42.7	45.5
STG-GCN [179]	10.1	16.9	32.5	38.5	50.0	-	-	72.9	10.0	11.9	20.1	24.0	30.4	-	-	43.1
siMLPe (Ours)	**4.5**	**9.8**	**22.0**	**28.1**	**39.3**	**49.2**	**57.8**	**63.7**	**6.1**	**10.8**	**19.1**	**22.8**	**29.5**	**35.1**	**39.7**	**42.7**

Note that the lengths T and N do not need to be equal. When $T > N$, we only take the N first frames of the prediction, and in the case of $T < N$, we could pad our input sequence to N by repeating the last frame, as done in [113, 111].

6.3.3 LOSSES

As mentioned in Section 6.1 and shown in Figure 6.1, the last input pose is "close" to the future poses. Inspired by this observation, instead of predicting the absolute 3D poses from scratch, we let our network predict the residual between the future pose P_{T+t} and the last input pose x_T. As we will show in Section 6.4.4, this eases learning and improves performance.

Objective function. Our objective function $\mathcal{L}$ includes two terms $\mathcal{L}_{re}$ and $\mathcal{L}_v$:

$$\mathcal{L} = \mathcal{L}_{re} + \mathcal{L}_v \ . \tag{6.7}$$

$\mathcal{L}_{re}$ aims to minimize the $\mathcal{L}_2$-norm between the predicted motion $\mathbf{P}_{T+1:T+N}$ and ground-truth one $\mathbf{G}_{T+1:T+N}$:

$$\mathcal{L}_{re} = \mathcal{L}_2(\mathbf{P}_{T+1:T+N}, \mathbf{G}_{T+1:T+N}) \ . \tag{6.8}$$

$\mathcal{L}_v$ aims to minimize the $\mathcal{L}_2$-norm between the velocity of the predicted motion $\mathbf{v^{GT}}_{T+1:T+N}$ and the ground truth one $\mathbf{v}_{T+1:T+N}$:

$$\mathcal{L}_v = \mathcal{L}_2(\mathbf{v^{GT}}_{T+1:T+N}, \mathbf{v}_{T+1:T+N}) \ , \tag{6.9}$$

where $\mathbf{v}_{T+1:T+N} = [v_{T+1}^\top, .., v_{T+N}^\top]^\top \in \mathbb{R}^{N \times C}$, v_t represents the velocity at frame t and is computed as the time difference: $v_t = P_{t+1} - P_t$. We provide a full analysis of the loss terms in Section 6.4.4.

6.4 EXPERIMENTS FOR SINGLE-PERSON HUMAN MOTION PREDICTION

In this section, we present our experimental details and results. We introduce the datasets and evaluation metric in Section 6.4.1, the implementation details in Section 6.4.2, and the quantitative/qualitative results in Section 6.4.3. An exhaustive ablation analysis is provided in Section 6.4.4.

6.4.1 DATASETS AND EVALUATION METRIC

Human3.6M dataset [75]. Human3.6M contains 7 actors performing 15 actions, and 32 joints are labeled for each pose. We follow the same testing protocols of [111] and

Table 6.5: **Ablation of the number of MLP blocks** on Human3.6M.

Nb. Blocks	# Param.(M) ↓	MPJPE (mm) ↓							
		80	160	320	400	560	720	880	1000
1	0.012	12.7	28.5	59.7	72.1	93.6	107.0	116.8	123.6
2	0.014	10.9	24.9	52.3	64.0	83.2	97.3	108.4	115.4
6	0.025	10.2	23.1	48.8	60.1	79.0	93.3	105.1	112.6
12	0.041	9.9	22.4	47.2	58.3	77.1	91.5	103.3	110.9
24	0.073	9.7	22.0	46.8	57.7	76.4	90.8	102.6	110.3
48 (Ours)	0.138	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**
64	0.180	9.6	21.8	46.5	57.5	76.0	90.1	101.9	109.7
96	0.266	9.7	21.9	46.7	57.8	76.3	90.5	102.1	109.8

Table 6.6: **Ablation of different components of our network** on Human3.6M. *'LN'* denotes the layer normalization. *'DCT'* denotes the DCT transformation. *'Spa. only'* means that all FC layers are on spatial dimensions (w/o transpose operations before/after MLP blocs). *'Temp. only'* means that all FC layers are on temporal dimensions (w/o any transpose operations).

Ablation	80	160	320	400	560	720	880	1000
Spa. only, w/o LN	23.7	44.0	75.5	87.6	106.3	120.4	130.5	135.6
Spa. only	23.8	43.0	73.4	85.2	102.0	116.3	125.3	131.9
Temp. only	9.9	22.4	47.2	58.4	77.2	91.1	102.8	110.5
w/o LN	12.7	29.0	62.3	76.2	97.4	111.6	121.6	127.3
w/o DCT	9.9	22.4	47.3	58.4	76.9	91.2	102.8	110.5
SIMLPE (ours)	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**

use *S5* as the test set, *S11* as the validation set, and the others as the train set. Previous works use different test sampling strategies, including 8 samples per action [114, 113], 256 samples per action [111] or all samples in the test set [41]. As 8 samples are too few and taking all testing samples could not balance different actions with different sequence lengths, we thus take 256 samples per action for testing and evaluate on 22 joints as in [114, 113, 111, 107].

Table 6.7: **Ablation of data augmentation** on Human3.6M. We only use front-back flip as our data augmentation, i.e., we randomly invert the motion sequence during the training.

	80	160	320	400	560	720	880	1000
w/o aug	10.0	22.6	48.3	59.7	78.2	92.0	103.4	110.8
w aug	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**

AMASS dataset [109]. AMASS is a collection of multiple Mocap datasets [48, 109, 12, 85, 155, 157, 21, 106, 51, 31, 143, 110, 104, 124, 3, 158, 71, 156] unified by SMPL parameterization [105]. We follow [111] to use AMASS-BMLrub [155] as the test set and split the rest of the AMASS dataset into training and validation sets. The model is evaluated on 18 joints as in [111].

3DPW dataset [163]. 3DPW is a dataset including indoor and outdoor scenes. A pose is represented by 26 joints, but we follow [111] and evaluate 18 joints using the model trained on AMASS to evaluate generalization.

Evaluation metric. We report the Mean Per Joint Position Error (MPJPE) on 3D joint coordinates, which is the most widely used metric for evaluating 3D pose errors. This metric calculates the average L2-norm across different joints between the prediction and ground-truth. Similar to previous works [113, 111, 41, 107], we ignore the global rotation and translation of the poses and keep the sampling rate as 25 frames per second (FPS) for all datasets.

6.4.2 IMPLEMENTATION DETAILS

In practice, we set the input length $T = 50$, the output length $N = 10$ on the Human3.6M dataset, and $N = 25$ on AMASS dataset and 3DPW dataset. During testing, we apply our model in an auto-regressive manner to generate motion for longer periods. The feature dimension $C = 3 \times \mathbf{J}$, where $\mathbf{J}$ is the number of joints, $\mathbf{J} = 22$ for Human3.6M and $\mathbf{J} = 18$ for AMASS and 3DPW.

To train our network, we set the batch size to 256 and use the Adam optimizer [78]. The memory consumed by our network is about $1.5GB$ during the training. All our experiments are conducted using the Pytorch [129] framework on a single NVIDIA RTX 2080Ti graphics card. We train our network on the Human3.6M dataset for 35k iterations, the learning rate starts from 0.0003 at the beginning and drops to 0.00001 after 30k steps. The training takes ∼30 minutes. For AMASS dataset, we train our network for 115k iterations. The learning rate starts from 0.0003 at the beginning and drops to 0.00001 after 100k steps. The training takes ∼2 hours. During training, we only use the front-back flip as data augmentation, which randomly inverts the motion sequence during the training.

6.4.3 QUANTITATIVE AND QUALITATIVE RESULTS

In this section, we compare our approach to existing state-of-the-art methods on different datasets. We report MPJPE in *mm* at different prediction time steps up to 1000ms.

Human3.6M dataset. In Table 6.1, we compare our method with other state-of-the-art methods on the Human3.6M dataset. Our method outperforms all previous methods on every frame with much fewer parameters.

As explained in Section 6.4.1, some different methods have taken different test sampling strategies. Following [111], we choose to test with 256 samples on 22 joints. To make a fair comparison, we evaluate all the methods using the same testing protocol. Our method outperforms all previous methods on every frame with a much less number of parameters. Besides, previous works usually report short-term ($0 \sim 500ms$) and long-term ($500 \sim 1000ms$) predictions separately, and [107] reports short-/long- term results using two different models. In our tables, all the results from $0 \sim 1000ms$ are predicted by a single model, and for [107], we report the results of their model which achieves the best performance on long-term prediction. In addition, we also evaluate the two simple approaches mentioned in Section 6.1 on the Human3.6M dataset in Table 6.1: ‘Repeating Last-frame’ takes the last input pose and repeats it N times to serve as output, and ‘One FC’ uses only one single fully connected layer trained on Human3.6m. These results show that the task of human motion prediction could be potentially modeled in a completely different and simple way without explicitly fusing spatial and temporal information. Furthermore, similar to all the previous works, we also detail the action-wise results in Table 6.2.

AMASS and 3DPW datasets. In Table 6.3, we report the performance of the model trained on AMASS and tested on the AMASS-BMLrub and 3DPW datasets, following the evaluation protocol of [111]. Different from the Human3.6M dataset where the training and testing data are from the same types of actions performed by different actors, the difference between training and testing data under this protocol is much larger, which makes the task more challenging in terms of generalization. As shown in the table, our approach performs consistently better on long-term prediction. Moreover, our model is much lighter. For example, the parameter size of our model is $\sim 4\%$ of Hisrep [111].

While the commonly used evaluation protocol is to consider the predicted error at different time steps, some works [146, 179] report their result by taking the average error from the first time step to a certain time step. We report the predicted error at different time steps in all the tables, except in Table 6.4, where we report the average error for comparison with [146, 179]. Our approach also achieves better performance than these two methods.

Qualitative results. In addition to the quantitative results, we provide some qualitative results of our method in Figure 6.3, showing some testing examples on the Human3.6M dataset. We could find that the predictions of our method perfectly match the ground-truth on short-term prediction, and globally fits the ground-truth on long-term prediction. The error becomes larger when looking into longer predictions, which is a common problem for all the motion prediction methods as shown in Table 6.1 and Table 6.3. This is because

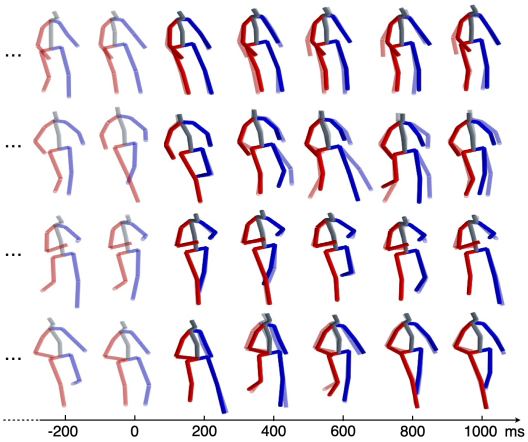


Figure 6.3: **Qualitative results of our method siMLPe.** The skeletons in light colors are the input (before 0ms) and the ground-truth (after 0ms). Those with dark colors represent the predicted motions. Our prediction results are close to the ground-truth.

most of the current methods use auto-regression for predicting a longer future, which will make the error accumulate. Moreover, uncertainty grows very quickly with time when predicting human motions.

6.4.4 ABLATION STUDY

We evaluate below the influence of the different components of our approach on the Human3.6M dataset.

Number of MLP blocks. We ablate the number of MLP blocks m in Table 6.5. Our proposed architecture already achieves good performance using only 2 MLP blocks with $0.014M$ parameters. The network achieves its best performance with 48 MLP blocks.

Network architecture. In Table 6.6, we ablate the different components of our network. As the table shows, temporal feature fusion and layer normalization are both of vital importance to our network. If the network just operates along the spatial dimension of the motion sequence without merging any information across different frames, it will lead to

Table 6.8: **Ablation of different loss terms** on Human3.6M.

$\mathcal{L}_{re}$	$\mathcal{L}_v$	80	160	320	400	560	720	880	1000
✓		9.6	21.8	46.5	57.5	76.7	91.5	103.5	111.3
✓	✓	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**

degraded results. However, if the network just operates along the temporal dimension, the network will still achieve comparable performance. Besides, the use of DCT transformation can further improve performance slightly.

Data augmentation. In Table 6.7, we ablate the use of front-back flip data augmentation and find that the data augmentation slightly improves the performance.

Loss. In Table 6.8, we evaluate the importance of different loss terms used during training. As shown in the table, with the help of the velocity loss $\mathcal{L}_v$, the network achieves better performance on long-term predictions while maintaining the same performance on the short-term.

Learning residual displacement. In Table 6.9, we analyze the importance of the proposed residual displacement and compare it to other types of residual used in previous works [114, 113]. Our method aims to predict the differences between each future pose and the last observed pose, after the IDCT transformation. When predicting directly the absolute 3D pose ('w/o residual'), the performance drops dramatically. We also test other types of residual by either learning the residual in the DCT space, before applying the IDCT transformation ('Before IDCT') following [113], or learning the velocity of the motion ('consecutive') following [114], and both achieve inferior performance compared to our proposed residual displacement.

Table 6.9: **Analysis of different types of residual displacement** on Human3.6M. SIMLPE predicts the differences of each future frame with the last observation (after IDCT). *'Before IDCT'* learns the residual before applying the IDCT transformation. *'Consecutive'* learns the velocity between consecutive frames. *'w/o residual'* predicts directly the absolute 3D poses.

Residual	80	160	320	400	560	720	880	1000
w/o residual	12.4	25.1	50.7	61.6	80.1	93.9	105.5	113.0
Consecutive	9.7	22.0	46.8	57.8	76.5	90.7	102.4	110.1
Before IDCT	10.4	23.0	48.2	59.1	77.9	91.8	103.2	110.5
SIMLPE (ours)	**9.6**	**21.7**	**46.3**	**57.3**	**75.7**	**90.1**	**101.8**	**109.4**

6.5 EXPERIMENTS FOR MULTI-PERSON HUMAN MOTION PREDICTION

6.5.1 DATASETS AND EVALUATION METRIC

We test our method on ExPI dataset [60]. ExPI contains 2 couple of actors performing 16 actions, and 18 joints are labeled for each actor. We follow the test protocols of [60] and test on the common action splits. To evaluate the performance of the proposed methods on ExPI, we evaluate our proposed method along with other methods on JME and AME introduced in Section 5.6.2.

6.5.2 QUANTITATIVE RESULTS

In order to implement SiMLPe in the multi-person scenario, we tried several different implementation strategies: "mix": we train a single SiMLPe model using data of the two people (Figure 6.4); "cat": we concatenate the two persons together as a single instance; "sep": we train two models separately for the two persons (Figure 6.5). Besides, we also designed a simple architecture with a link bloc between the two branches for the two people. The link bloc takes the embeddings of the two branches and concatenates them as input to an FC layer, and then we add the output back to the two branches (Figure 6.6).

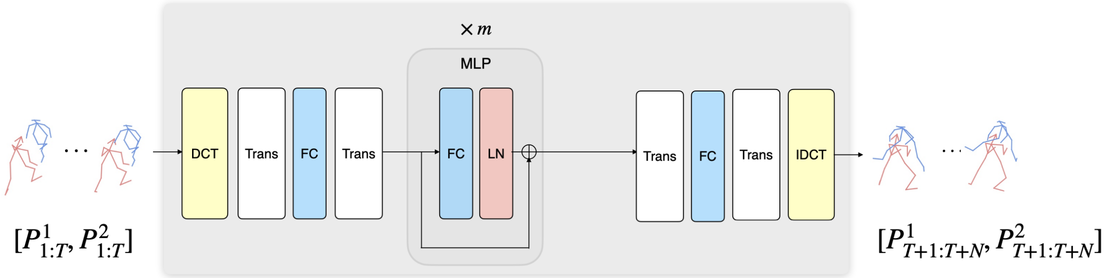


Figure 6.4: Pipeline of SIMLPE on ExPI dataset by concatenating two persons together as a single instance (“cat”).

Figure 6.8 and Figure 6.7 show experimental results of SiMLPe and previous SOTA methods on ExPI datasets. The red colors show methods designed for single-person scenarios and implemented on ExPI. The blue colors show methods designed for collaborative motion prediction with a link bloc between the two branches. We could observe that SiMPLe achieves comparable or even better results to the state of the art, and more importantly, the model size is much smaller than the previous SOTA. Besides, adding the link bloc for SiMLPe helps improve the performance of SIMLPE on the multi-person implementation.

6.6 CONCLUSION

In this chapter, we present SIMLPE, a simple yet effective network for human motion prediction. SIMLPE is composed of only fully connected layers, layer normalization, and transpose operations. The only non-linear operation is thus the layer normalization. While using much fewer parameters, SIMLPE achieves state-of-the-art performance on various benchmarks. The reported ablation study also demonstrates an interest in various design choices, highlighting the importance of temporal information fusion in this task. We hope the simplicity of SIMLPE will help the community to rethink the task of human motion prediction.

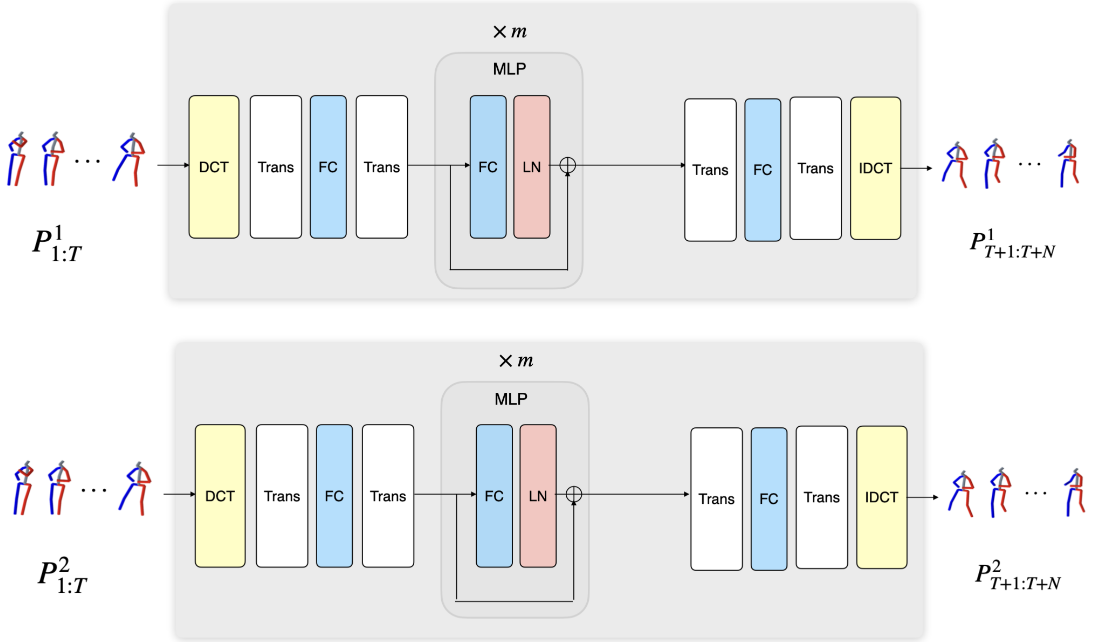

$\times m$
MLP
DCT
Trans
FC
Trans
FC
LN
Trans
FC
Trans
IDCT
$P^1_{1:T}$
$P^1_{T+1:T+N}$
$\times m$
MLP
DCT
Trans
FC
Trans
FC
LN
Trans
FC
Trans
IDCT
$P^2_{1:T}$
$P^2_{T+1:T+N}$

Figure 6.5: Pipeline of SIMLPE on ExPI dataset by training two models separately for the two persons ("sep").

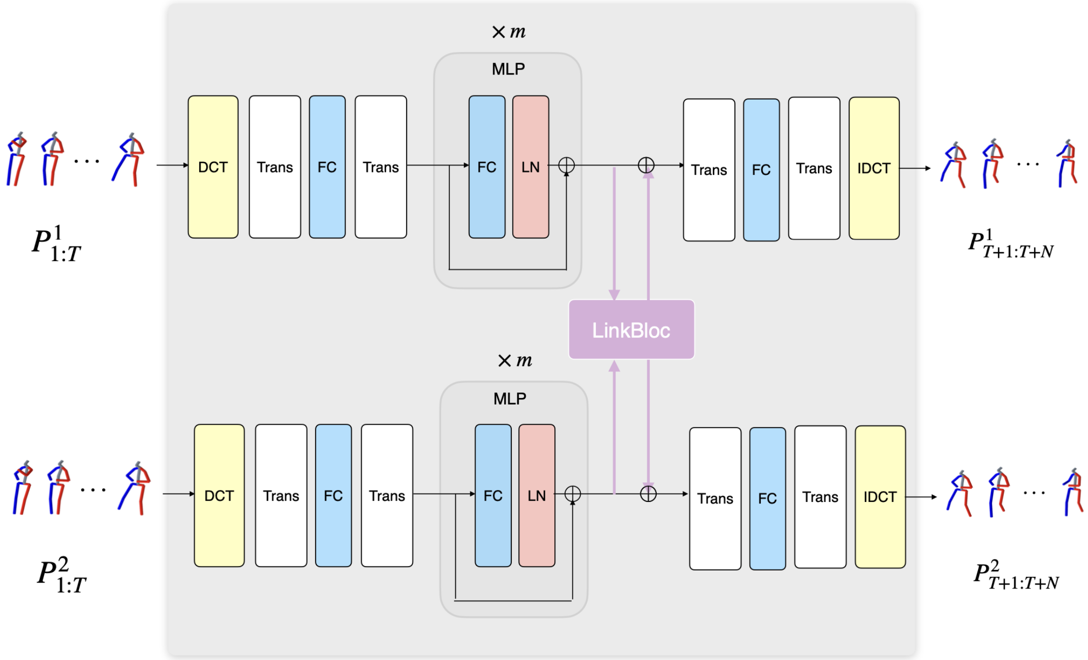

$P^1_{1:T}$
DCT
Trans
FC
Trans
× m
MLP
FC
LN
LinkBloc
Trans
FC
Trans
IDCT
$P^1_{T+1:T+N}$
$P^2_{1:T}$
DCT
Trans
FC
Trans
× m
MLP
FC
LN
Trans
FC
Trans
IDCT
$P^2_{T+1:T+N}$

Figure 6.6: Pipeline of SIMLPE on ExPI dataset by adding a link bloc between the two branches (“link”).

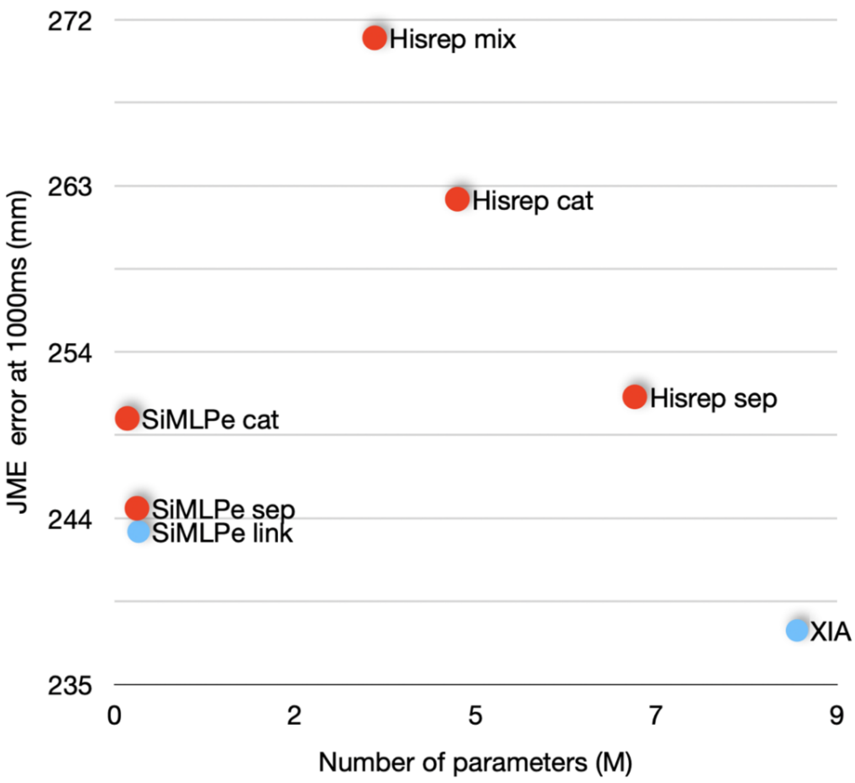


Figure 6.7: Comparison of parameter size and JME performance on the ExPI dataset. We report the JME metric in mm at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. The red colors show methods designed for single-person scenarios and implemented on ExPI. The blue colors show methods designed for collaborative motion prediction with a link bloc between the two branches. XIA is the sota on ExPI which is introduced in Section 5.

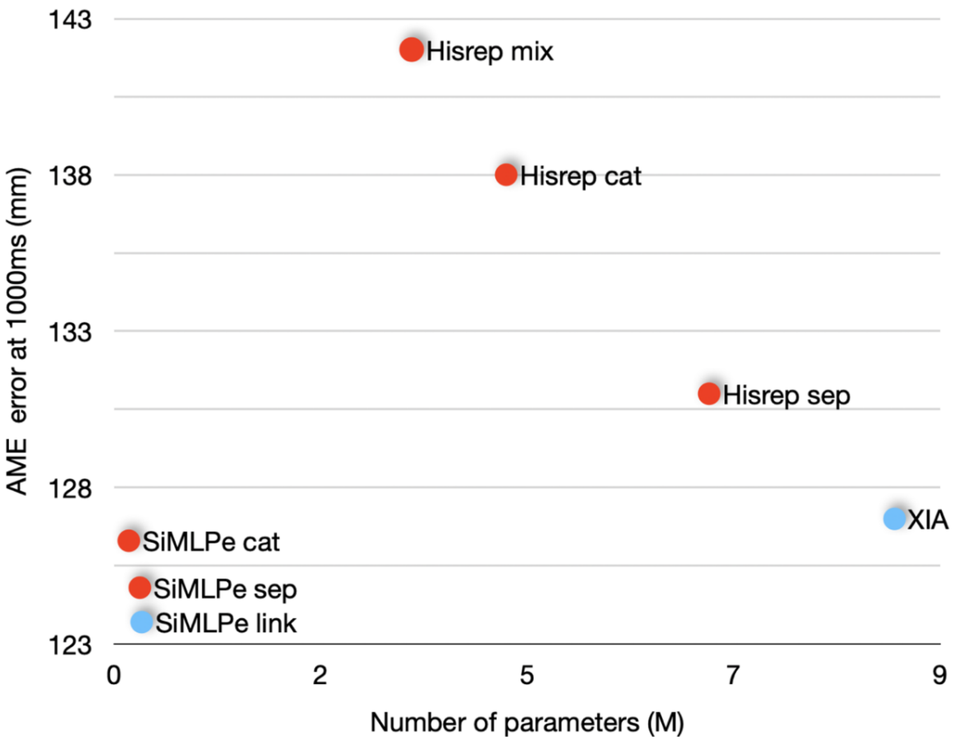


Figure 6.8: Comparison of parameter size and AME performance on the ExPI dataset. We report the AME metric in mm at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. The red colors show methods designed for single-person scenarios and implemented on ExPI. The blue colors show methods designed for collaborative motion prediction with a link bloc between the two branches. XIA is the sota on ExPI which is introduced in Section 5.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this chapter, we summarize the main contributions of this thesis and discuss some research directions for future work.

7.1 SUMMARY

In this thesis, we focus on studying human poses and motions, considering the multi-person scenario where people are interacting. We aim to explore the use of interaction among different people to improve the understanding of human pose and motion. Based on this, we have focused on two tasks, human pose estimation from a single frame, and human motion prediction at the sequence level.

In Chapter 4, we started by studying human pose estimation from a single-frame RGB image. We take the raw predictions of multi-person from a state-of-the-art predictor that treats each instance separately and then use a GRU-based network named PI-net to model and learn the interaction relations among different people. Our refined poses outperform the raw ones which verifies the feasibility of studying person interaction.

Due to the irregularity and uncertainty of human motion (usually we would misunderstand an action by just looking into one single frame, for example, if two people are standing still face to face, we cannot know if they are talking to each other or just passing by), we continue to study the sequence level of human poses to better understand human interaction, and thus in Chapter 5 and Chapter 6, we focus on the task of human motion prediction.

In Chapter 5, we proposed collaborative motion prediction which aims at predicting the future motion of highly interacted people from an observed motion sequence. We also proposed a cross-interaction network with the guidance of the interacted person, which outperforms the previous state-of-the-art methods designed for a single person. To enable the study of this problem, we also collected a new dataset named ExPI which contained highly interacted and challenging poses suitable for the study of multiple human pose understanding tasks.

In Chapter 6, we propose siMLPe, a simple yet effective network for human motion prediction that is based on MLP layers and only has layer normalization as no-linear components. The proposed model is about 30 times smaller than the state-of-the-art network but surpasses the previous state-of-the-art work on multiple benchmarks. This brings some new insights into the design of networks in the human motion prediction task, as well as other human motion-related tasks.

7.2 FUTURE RESEARCH DIRECTIONS AND DISCUSSIONS

From this thesis, several lines of future work could be explored for understanding 3D human pose and motion. In this section, we will firstly discuss some possible direct follow-up works based on the works introduced in this these, and then discuss some other related research directions within the domain in a broader view.

7.2.1 FUTURE FOLLOW-UP WORKS

Based on the above works discussed in this thesis, there are several possible follow-up directions closely related:

- Application of MLP-based architecture on other tasks of human motion: In Chapter 6, we proposed a simple but strong baseline network that achieves state-of-the-art performance on the task of human motion prediction. Some recent literature has also shown the effectiveness of this network on other human motion tasks such as generating whole-body motion from certain joints [45]. Thus we are working on extending the MLP-based simple architecture to more motion tasks such as single and multi-motion generation from a sequence of motion or from speech and audio.
- Represent interaction explicitly: As discussed at the beginning of this thesis in Chapter 1, one challenge for studying person interaction is to think about how to model and learn the interaction information. The works presented in this thesis use networks to embed and learn the interaction information implicitly, which has proven to be an effective solution, and some recent works also follow a similar manner [141].

However, another possibility is to model interactions explicitly. There are two main motivations for this approach: Firstly, as mentioned in Chapter 3, an individual can be represented by a kinematic tree, which models the structural information and links between different joints. However, it is difficult to apply a similar approach to multi-person scenarios, making it reasonable to explore alternative representations that achieve similar objectives. Secondly, with the success of SiMLPe, we have learned that a simple network can effectively learn features for motion data. Instead of designing a network specifically for learning implicit interactions, we might be able to use a simple network to learn the temporal dependencies of motion data while focusing on input data formulation to explicitly model interactions. If successful, this approach could be generalized to model various types of interactions, such as those between people and objects, or consider person-person and person-object interactions simultaneously.

- ”Dancing with your partner”: In Chapter 5,we introduced collaborative motion prediction, where we predict the future motion sequence of two individuals based on their previous observations. In highly interactive scenarios, such as dancing, where two people are closely connected, it would be interesting to study the possibility of inferring the motion of one person based on the observation of the other. This would allow the generated character to dance in sync with the given partner, further enhancing the understanding of human interactions in such dynamic environments.

7.2.2 A BROADER DISCUSSION WITHIN THE DOMAIN OF 3D HUMAN POSE AND MOTION UNDERSTANDING

In addition to the direct follow-up work discussed above, we will discuss some other research directions that are interesting and closely related to this thesis.

3D human understanding with other context information. In the real world, a person always interacts with the surroundings, thus considering context information in the understanding of humans is important. Considering 3D human understanding, there are three main types of context information:

- Person-person interaction, which models the interaction of different persons under the same scene, especially involved in the same activity;
- Person-object interaction, which models the interaction of a person with the surrounded or contacted objects.
- Physical constraints, which studies the human pose under the consideration of physical constraints such as forces.

In this thesis, we only focused on the first type of person-person interaction, while the other two types of context are also starting to attract more attention in recent years [177, 18, 63, 73, 64]. Digging into all these kinds of context information, or combining different kinds of context information together will no doubt be an important direction in the future of 3D human understanding.

Data-lacking problem in 3D motion tasks. Research in 2D computer vision tasks has developed very quickly in the past years with the help of deep learning methods. As a data-driven problem, one important support for quick development is the collection of large-scale datasets [101, 10]. With enough human and funding resources, huge-scale 2D datasets could be labeled and used for training big models. Unfortunately, this becomes more complicated in 3D vision. Labeling 3D data is more difficult than 2D as the collection and annotating are highly limited by the labeling equipment and actors. Though many 3D datasets [118, 75, 109, 163, 47, 59] are collected and released in recent years, the amount of data is still far from the size of 2D datasets. Besides, due to the limitation of the collecting process, data in current lab-based 3D datasets are not diverse enough. The public 3D human datasets usually contain a limited number of actions, and the background environment is also not varied enough.

Looking into the long-term development of 3D vision, one of the significant challenges that needs to be addressed is the problem of data scarcity. In addition to labeling massive

3D datasets, virtual datasets [159] and generated 3D data [108] can provide a complementary solution. Another potential solution is to explore unsupervised learning methods which is still a relatively underexplored area of research.

3D motion generation. As said in the above paragraph, generating human motion would be possible to help solve the data lacking problem in 3D human understanding. In addition to this potential application, human motion generation itself also has more meanings far more than this. Different from human motion prediction discussed in this thesis, which is a deterministic task aiming to predict the most possible future close to the ground truth, the task of human motion generation is a stochastic process: it focuses on generating various possibilities of the future to model the uncertainty of motion. The human motion could be generated from various information such as a given action labels [58, 133], text information [152, 134, 153], a sequence of music[97, 144], or an observed motion sequence [174, 137, 19]. These tasks have received increasing attention recently, while most of them focus on a single person. Some very recent works have started to study multi-person generation [141] but the problem is still not well explored. We believe that the studies and contributions in this thesis will also have the potential to help with the multi-person generation problems.

LIST OF FIGURES

1.1 Applications of human understanding: (a) sport analyses; (b) a hospitalization robotic which takes care of patients; (c) an entertainment robotic which could follow the motion of the user; (d) autonomous driving; (f) virtual avatar dancing with the user. [1] 14

3.1 Datasets used in this thesis. (a, b, c) are datasets we use for human pose estimation in Chapter 4, (d, e, f) are datasets we use for human motion prediction in Chapter 6, and (g) is a dataset we collect and present in Chapter 5. 32

3.2 Inria Kinovis-platform. 34

4.1 **PI-Net peformance.** An example of test results on MuPoTS dataset. Poses refined by PI-Net (in green) are closer to the ground truth (in black) than the baseline (in red). We zoom in to several parts to clearly appreciate the difference. The error before and after PI-Net refinement for each person is shown in the table. The average 3D joint error for this example is reduced from 88.02 mm to 86.19 mm. 38

4.2 **PI-Net Architecture**. Mask-RCNN [66] and PoseNet [122] are used to extract the initial pose estimates $\mathbf{P}_1, \ldots, \mathbf{P}_N$. These estimates are fed into PI-Net, composed of three main blocks: Bi-RNN, Self-attention, and the shared fully-connected layers. The output of PI-Net refines the initial pose estimates by exploiting the pose of the interactions, yielding $\mathbf{Q}_1, \ldots, \mathbf{Q}_N$. 43

4.3 **Qualitative results on the COCO dataset**. For each pose, a darker color is used to represent the left side of the person. The bottom-right example corresponds to a failure case, as the 'red' and 'black' persons should be located in front of the scene, behind the 'blue' and 'purple' persons. This is caused by a misdetection on the root position of the input detected poses provided by the baseline network, while our network designed for refining the poses could not refine this kind of large deviation because this large deviation caused by the baseline network hinders our PI-net from learning the correct context information for correctly interpreting and refining the prediction. 52

5.1 **Collaborative human motion prediction. 1st row:** 3D sample meshes from our ExPI Dataset (just for visualization purposes). **2nd-4th rows:** Motion prediction results by MSR [42], Hisrep [111], and our method. Dark red/blue indicate prediction results, and light red/blue are the ground truth. By exploiting the interaction information, our approach of collaborative motion prediction achieves significantly better results than methods that independently predict the motion of each person. 58

5.2 Some samples of the ExPI dataset: RGB image with projected 2D skeletons, 3D pose, mesh, and textured mesh. 61

5.3 An illustration of labeling the missing joints[2]. 66

5.4 Data-cleaning. **Top:**Data before cleaning. The two joints 'F-back' and 'F-fhead' are missed. **Bottom:** Data after cleaning. The yellow marks indicate the two relabeled joints. 67

5.5 Per-action extremeness comparison of Human3.6M dataset and ExPI dataset. This figure shows the percentage of joints whose *std* is among a certain threshold (in different colors), for different actions. Actions with more red colors are more extreme. 70

5.6 Average extremeness comparison of Human3.6M dataset and ExPI dataset. This figure shows the percentage of joints whose *std* is beyond a certain threshold. 71

5.7 Quantitative tests of multiple tasks on ExPI dataset: (a) 2D pose estimation by Openpose [29]; (b) Person detection by Mask-rcnn [66]; (c) Instance segmentation by Mask-rcnn [66]; (d) Pose estimation by Dense pose [57]; (e) 3D pose reconstruction by SPIN [81]; (f) Human shape estimation by VIBE [80] 72

5.8 Computing flow of the proposed method. Two parallel pipelines – for the leader and the follower – are implemented. The key-value pairs are refined by XIA modules(we just visualize XIA modules for the first sub-sequences, while it is the same for the following sub-sequences). 73

5.9 Cross-interaction attention (XIA) module. In order to refine w with the help of the corresponding interaction information $w_{int.}$, the multi-head attention is queried by $w_{int.}$ and take w as key and value. 74

5.10 (a): Percentages of improvement of our method comparing with different state-of-the-art methods, measured by average JME error on the common action split, at different forecast times. Lower value means closer performance with our model. Our method surpasses these methods up to $10 \sim 40\%$ in the short term, and $5 \sim 30\%$ in the long term. (b): Joint-wise JME improvement(mm) of our method over Hisrep [111] and MSR [95]. Darker color means larger improvement. 76

5.11 Visualization results of our proposed method compared with previous state-of-the-art methods. Dark red/blue shows predicted results, and light red/blue represents groundtruth. 86

5.12 Visualization results of our proposed method compared with previous state-of-the-art methods (continue). Dark red/blue shows predicted results, and light red/blue represents groundtruth. 87

5.13 Visualization results of our proposed method compared with previous state-of-the-art methods (continue). Dark red/blue shows predicted results, and light red/blue represents groundtruth. 88

6.1 **Comparison of parameter size and performance on the Human3.6M dataset [75].** We report the MPJPE metric in *mm* at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. Our method (SIMLPE, in red) achieves the lowest error with significantly fewer parameters. We also show the performance of two simple methods: 'Repeating Last-Frame' systematically repeats the last input frame as output prediction, and 'One-FC' uses only one single fully connected layer to predict the future motion. 91

6.2 **Overview of our approach SIMLPE for human motion prediction.** *FC* denotes a fully connected layer, *LN* denotes layer normalization [13], and *Trans* represents the transpose operation. *DCT* and *IDCT* represent the discrete cosine transformation and inverse discrete cosine transformations respectively. The MLP blocks (in gray), composing FC and LN, are repeated m times. 93

6.3 **Qualitative results of our method SIMLPE.** The skeletons in light colors are the input (before 0ms) and the ground-truth (after 0ms). Those with dark colors represent the predicted motions. Our prediction results are close to the ground-truth. 104

6.4 Pipeline of SIMLPE on ExPI dataset by concatenating two persons together as a single instance ("cat"). 107

6.5 Pipeline of SIMLPE on ExPI dataset by training two models separately for the two persons ("sep"). 108

6.6 Pipeline of SIMLPE on ExPI dataset by adding a link bloc between the two branches ("link"). 109

6.7 Comparison of parameter size and JME performance on the ExPI dataset. We report the JME metric in mm at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. The red colors show methods designed for single-person scenarios and implemented on ExPI. The blue colors show methods designed for collaborative motion prediction with a link bloc between the two branches. XIA is the sota on ExPI which is introduced in Section 5. 110

6.8 Comparison of parameter size and AME performance on the ExPI dataset. We report the AME metric in mm at 1000 ms as performance on the vertical axis. The closer to the bottom-left, the better. The red colors show methods designed for single-person scenarios and implemented on ExPI. The blue colors show methods designed for collaborative motion prediction with a link bloc between the two branches. XIA is the sota on ExPI which is introduced in Section 5. 111

LIST OF TABLES

4.1 Sequence-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, the fourth method and our model are tested with ground truth bounding boxes and roots. Higher value means better performance. 50

4.2 PA MPJPE (top) and MPJPE (bottom) comparisons of PI-net with the state-of-the-art method [122] used as our baseline on the MuPoTS dataset. The average value indicated the image-wise average. Ground truth bounding boxes and roots are used for testing. Lower value means better performance. 51

4.3 Joint-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, and the fourth method and our model are tested with ground truth bounding boxes and roots. All ground truths are used for evaluation. Higher value means better performance. 53

4.4 Comparison of different input orders. *Intuitive* is the one described in Section 4.3.2, from near to far. *Inverse* is the opposite. *Random* means in random order. 53

4.5 Importance of self-attention and bidirectionality (RNN). PI-Net uses a bidirectional RNN followed by a self-attention layer. We evaluate the impact of each of these choices: w/o Att. when removing attention, w/o Bi. considering standard RNN. 53

4.6 Ablating the unit of the interaction network: None [122], Graph Convolutional Networks (GCN); LSTM and Gated Recursive Units (GRU), with (2,3,4) layers. 54

5.1 Composition of the ExPI Dataset. The seven first actions are performed by both couples. Six more actions are performed by Couple 1, while three others by Couple 2. 68

5.2 Comparison of ExPI with other publicly available datasets commonly used for human motion prediction. 69

5.3 Results on common action split with the two evaluation metrics (in mm). Lower value means better performance. Obviously, our proposal outperforms all the other methods both on JME and AME. 80

5.4 Results on single action split with the two evaluation metrics (in mm). Lower value means better performance. Seven action-wise models are trained independently. Our method performs the best in 5 actions, and close to the best for the other 2 actions. 81

5.5 Action-wise results on unseen action split with the two evaluation metrics (in mm). Lower value means better performance. Our method still performs the best on most of the unseen actions and on the average result. 83

5.6 Ablations. *'mix /cat /sep'* use the single person motion prediction model (Hisrep [111]) for multi-person by: mixing two poses together / concatenate two poses as a single vector / train two person-specific models. *'w.o. XIA'* indicates training leader and follower in parallel using our defined loss without XIA module; *'XIA kqv / kq / kv / v'* use XIA module to update key, value and query of the temporal attention, or just some of them. 84

6.1 **Results on Human3.6M** for different prediction time steps (ms). We report the MPJPE error in *mm* and number of parameters (M) for each method. Lower is better. 256 samples are tested for each action. † indicates that the results are taken from the paper [111], ⋆ indicated that the results are taken from the paper [107]. Note that ST-DGCN [107] use two different models to evaluate their short-/long- term performance, here we report their results of a single model which performs better on long-term for fair comparison. We also show results of two simple baselines: 'Repeating Last-Frame' repeats the last input frame 25 times as output, 'One FC' uses only one single fully connected layer for the prediction. 95

6.2 **Action-wise results on Human3.6M** for different prediction time steps (ms). Lower is better. 256 samples are tested for each action. † indicates that the results are taken from the paper [111], ⋆ indicates that the results are taken from the paper [107]. 96

6.3 **Results on AMASS and 3DPW** for different prediction time steps (ms). We report the MPJPE error in *mm*. Lower is better. The model is trained on the AMASS dataset. The results of the previous methods are taken from [111]. 98

6.4 **Average results for different prediction time periods on Human3.6M and AMASS.** These results are obtained following the evaluation method of STS-GCN [146] and STG-GCN [179], instead of the standard evaluation protocol adopted in [113, 111, 107]. 98

6.5 **Ablation of the number of MLP blocks** on Human3.6M. 100

6.6 **Ablation of different components of our network** on Human3.6M. *'LN'* denotes the layer normalization. *'DCT'* denotes the DCT transformation. *'Spa. only'* means that all FC layers are on spatial dimensions (w/o transpose operations before/after MLP blocs). *'Temp. only'* means that all FC layers are on temporal dimensions (w/o any transpose operations). 100

6.7 **Ablation of data augmentation** on Human3.6M. We only use front-back flip as our data augmentation, i.e., we randomly invert the motion sequence during the training. 101

6.8 **Ablation of different loss terms** on Human3.6M. 105

6.9 **Analysis of different types of residual displacement** on Human3.6M. siMLPe predicts the differences of each future frame with the last observation (after IDCT). *'Before IDCT'* learns the residual before applying the IDCT transformation. *'Consecutive'* learns the velocity between consecutive frames. *'w/o residual'* predicts directly the absolute 3D poses. 106

BIBLIOGRAPHY

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020.

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13390–13400, October 2021.

[3] Ijaz Akhter and Michael J. Black. Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction. In *CVPR*, 2015.

[4] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *arXiv e-prints*, pages arXiv–2004, 2020.

[5] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A Spatio-Temporal Transformer for 3D Human Motion Prediction. In *3DV*, 2021.

[6] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[7] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021.

[8] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020.

[9] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018.

[10] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele.

[11] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2014.

[12] Andreas Aristidou, Ariel Shamir, and Yiorgos Chrysanthou. Digital Dance Ethnography: Organizing Large Dance Collections. *J. Comput. Cult. Herit.*, 12(4), November 2019.

[13] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. 2016.

[14] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7412–7420, 2019.

[15] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human

motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018.

[16] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6856–6865, 2020.

[17] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 584–588. IEEE, 2019.

[18] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[19] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. 2022.

[20] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[21] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering Human Bodies in Motion. In *CVPR*, 2017.

[22] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. *arXiv preprint arXiv:2207.00499*, 2022.

[23] Matthew Brand and Aaron Hertzmann. Style Machines. In *Computer Graphics and Interactive Techniques*, 2000.

[24] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.

[25] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4563–4570. IEEE, 2018.

[26] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020.

[27] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021.

[28] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[29] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[30] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d

pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[31] Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, Shishir Subramanyam, Bart Kevelham, Caecilia Charbonnier, Pablo Cesar, Dimitrios Zarpalas, and Others. HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media. 2020.

[32] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[33] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[35] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2262–2271, 2019.

[36] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020.

[37] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[38] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[39] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2020.

[40] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)*, pages 405–414. IEEE, 2019.

[41] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *ICCV*, 2021.

[42] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021.

[43] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009.

[44] Yuming Du, Wen Guo, Yang Xiao, and Vincent Lepetit. 1st place solution for the uvo challenge on image-based open-world segmentation 2021. *arXiv preprint arXiv:2110.10239*, 2021.

[45] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Art-

siom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023.

[46] Panna Felsen, Patrick Lucey, and Sujoy Ganguly. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 732–747, 2018.

[47] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020.

[48] Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset.

[49] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[50] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[51] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A Large Multipurpose Motion and Video Dataset, 2020.

[52] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[53] Haifeng Gong, Jack Sim, Maxim Likhachev, and Jianbo Shi. Multi-Hypothesis Motion Planning for Visual Object Tracking. In *ICCV*, 2011.

[54] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[55] CMU graphics lab. Cmu graphics lab motion capture data- base., 2009. `http://mocap.cs.cmu.edu//`.

[56] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018.

[57] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[58] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. pages 2021–2029, 2020.

[59] Wen Guo. Multi-person pose estimation in complex physical interactions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4752–4755, 2020.

[60] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-Person Extreme Motion Prediction. In *CVPR*, 2022.

[61] Wen Guo, Enric Corona, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Pi-net: Pose interacting network for multi-person monocular 3d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2796–2806, 2021.

[62] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

[63] Vladimir Guzov, Torsten Sattler, and Gerard Pons-Moll. Visually plausible human-object interaction capture from wearable sensors. In *arXiv*, May 2022.

[64] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11354–11364, Piscataway, NJ, October 2021. IEEE.

[65] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019.

[66] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[68] Alejandro Hernandez, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. 2019.

[69] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.

[70] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015.

[71] Ludovic Hoyet, Kenneth Ryall, Rachel Mcdonnell, and Carol O'sullivan. Sleight of Hand: Perception of Finger Motion from Reduced Marker Sets. 2012.

[72] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[73] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022.

[74] Yingfan Huang, HuiKun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281, 2019.

[75] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[76] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

[77] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ARXIV*, 2014.

[79] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[80] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[81] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[82] Hema Swetha Koppula and Ashutosh Saxena. Anticipating Human Activities for Reactive Robotic Response. In *IROS*, 2013.

[83] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32:137–146, 2019.

[84] Laxman Kumarapu and Prerana Mukherjee. Animepose: Multi-person 3d pose estimation and animation. *arXiv preprint arXiv:2002.02792*, 2020.

[85] Bio Motion Lab. BMLhandball Motion Capture Database.

[86] Tim Lebailly, Sena Kiciroglu, Mathieu Salzmann, Pascal Fua, and Wei Wang. Motion Prediction Using Temporal Inception Module. 2020.

[87] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient Nonlinear Markov Models for Human Motion. In *CVPR*, 2014.

[88] Anliang Li, Shuang Wang, Wenzhu Li, Shengnan Liu, and Siyuan Zhang. Predicting human mobility with federated learning. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 441–444, 2020.

[89] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[90] Jiachen Li, Hengbo Ma, Zhihao Zhang, Jinning Li, and Masayoshi Tomizuka. Spatio-temporal graph dual-attention network for multi-agent prediction and tracking. *arXiv preprint arXiv:2102.09117*, 2021.

[91] Jiachen Li, Fan Yang, Hengbo Ma, Srikanth Malla, Masayoshi Tomizuka, and Chiho Choi. Rain: Reinforced hybrid attention inference network for motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16096–16106, October 2021.

[92] Jiefeng Li, Can Wang, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. *arXiv preprint arXiv:2008.00206*, 2020.

[93] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[94] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic Graph Neural Networks for 3D Skeleton-Based Human Action Recognition and Motion Prediction. 2021.

[95] Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 854–864, October 2021.

[96] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *CVPR*, 2020.

[97] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021.

[98] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[99] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[100] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. *arXiv preprint arXiv:1811.08264*, 2018.

[101] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[102] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[103] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019.

[104] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion And Shape Capture from Sparse Markers. 33(6), November 2014.

[105] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. 34(6), 2015.

[106] Eyes JAPAN Co Ltd. Eyes Japan MoCap Dataset.

[107] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *CVPR*, 2022.

[108] Takahiro Maeda and Norimichi Ukita. Motionaug: Augmentation with physical correction for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6427–6436, 2022.

[109] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019.

[110] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT Whole-Body Human Motion Database. In *ICAR*, 2015.

[111] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

[112] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021.

[113] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[114] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[115] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[116] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the

wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[117] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019.

[118] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[119] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018.

[120] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[121] Terry Monaghan. Why study the lindy hop? *Dance Research Journal*, 33(2):124–127, 2001.

[122] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10133–10142, 2019.

[123] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[124] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation Mocap Database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[125] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. 2010.

[126] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, pages 2277–2287, 2017.

[127] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

[128] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles. 2016.

[129] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *NIPS*, 2019.

[130] Lourenço V Pato, Renato Negrinho, and Pedro MQ Aguiar. Seeing without looking: Contextual rescoring of object detections for ap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14610–14618, 2020.

[131] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[132] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7025–7034, 2017.

[133] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, October 2021.

[134] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. 2022.

[135] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[136] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4342–4351, 2019.

[137] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. pages 11488–11499, 2021.

[138] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[139] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[140] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. Instance- and category-level 6d object pose estimation. In *RGB-D Image Analysis and Processing*, pages 243–265. Springer, 2019.

[141] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[142] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[143] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[144] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022.

[145] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11209–11218, October 2021.

[146] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-Time-Separable Graph Convolutional Network for Pose Forecasting. In *ICCV*, 2021.

[147] Alessandro Sperduti and Antonina Starita. Supervised Neural Networks for the Classification of Structures. 1997.

[148] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[149] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[150] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pages 529–545, 2018.

[151] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling Human Motion Using Binary Latent Variables. In *NIPS*, 2007.

[152] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. 2022.

[153] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[154] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[155] Nikolaus F. Troje. Decomposing Biological Motion: A Framework for Analysis and Synthesis of Human Gait Patterns. *Journal of Vision*, 2(5), September 2002.

[156] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *BMVC*, 2017.

[157] Carnegie Mellon University. CMU MoCap Dataset.

[158] Simon Fraser University and National University of Singapore. SFU Motion Capture Database.

[159] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[160] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[161] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[162] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 4601–4607. IEEE, 2018.

[163] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[164] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision*, pages 601–617, 2018.

[165] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian Process Dynamical Models. In *NIPS*, 2005.

[166] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian Process Dynamical Models for Human Motion. 2007.

[167] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019.

[168] Xiaolin K Wei and Jinxiang Chai. Modeling 3d human poses from uncalibrated monocular images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1873–1880. IEEE, 2009.

[169] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. *arXiv preprint arXiv:2008.09457*, 2020.

[170] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019.

[171] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision*, pages 466–481, 2018.

[172] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[173] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[174] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.

[175] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of mul-

tiple scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[176] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, pages 8410–8419, 2018.

[177] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022.

[178] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[179] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-Temporal Gating-Adjacency GCN For Human Motion Prediction. In *CVPR*, 2022.

[180] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.